Bioinformatics approaches to identify pain mediators, novel LncRNAs and distinct modalities of neuropathic pain

by

Georgios Baskozos

A thesis submitted to

University College London

for the degree of

Doctor of Philosophy

Institute of Structural and Molecular Biology

University College London

September 2016

Declaration

I, Georgios Baskozos, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

……

Georgios Baskozos

29 September 2016

Abstract

This thesis presents a number of studies in the general subject of bioinformatics and functional genomics. The studies were made in collaboration with experimental scientists of the London Pain Consortium (LPC), an initiative that has promoted collaborations between experimental and computational scientists to further understanding of pain. The studies are mainly concerned with the molecular biology of pain and deal with data gathered from high throughput technologies aiming to assess the transcriptional changes involved in well induced pain states, both from animal models of pain and human patients. We have analysed next generation sequencing data (NGS data) in order to assess the transcriptional changes in rodent's dorsal root ganglions under well induced pain states. We have also developed a customised computational pipeline to analyse RNA-sequencing data in order to identify novel Long non-coding RNAs (LncRNAs), which may function as mediators of neuropathic pain. Our analyses detected hundreds of novel LncRNAs significantly dysregulated between sham-operated animals and animal models of pain. In addition, in order to gain valuable insights into neuropathic pain, including both its molecular signature, somatosensory profiles and clusters of individuals related to pain severity, we analysed clinical data together with data obtained from quality of life pain-questionnaires. Based on this study, we were able to identify distinct pain modalities associated with the intensity of neuropathic pain. Our results will be useful for the understanding of neuropathic pain and its future treatment.

Acknowledgments

I would like to thank my supervisor Christine Orengo for all her irreplaceable guidance, help and support throughout this PhD and for giving me the opportunity to work in such a friendly and prestigious group. I would like to thank David Bennett for all the immeasurable guidance and support throughout my research and for giving me the opportunity to collaborate with such prestigious scientists and groups. Their help and advise throughout these years have been really immeasurable. I would also like to thank Steve MacMahon and all scientists of the London Pain Consortium for their advice and for educating me about pain and the nervous system. I would also like to acknowledge my subsidiary supervisor Andrew Martin and chair of the thesis committee Kevin Bryson for all the help and guidance.

This work has been done in collaboration with many groups. First I would like to thank all members of the Orengo group, past and present, for all their advice and help and for creating the friendly and supportive lab where I have always been enjoying working in. Also I would like to thank all the members of David Bennett's group and Steve MacMahon’s group. In particular I would like to thank Jim Perkins, Ana Antunes-Martins, John Dawes, John Lees and Andreas Themistocleous. It has always been a pleasure to work with them. Many thanks also to Jeffrey Mogil and all people in his group, with whom I collaborated and they have been really hospitable and supportive.

Finally, a special thank to all my friends and family, both here in the UK and back in Greece. Thank you for being there and gave me the courage to take this path. This PhD would not have been possible without your support.

Contents

Table of Contents

Abstract....3
Acknowledgments....4
Contents....5
List of Figures....8
List of tables....11
Introduction....12
- Pain....13
  - Pain at the molecular level....15
  - Neuropathic Pain....20
  - Animal models of pain....22
- Gene Expression....26
  - The Central Dogma....26
  - Long Non-coding RNAs (LncRNAs)....27
  - Functional repertoires of LncRNAs....35
  - Known pain-related LncRNAs....37
  - Overview of computational pipelines for identifying LncRNAs…38
- RNA-Sequencing....39
  - RNA isolation and library construction....41
 44
  - Potentials and drawbacks....44
  - Analysing RNA-sequencing data....47
- Explain complex interactions of many variables....55
  - Principal Components Analysis and varimax rotation....55
- Overview of thesis chapters....57

Methods for identifying LncRNAs and analyse RNA-sequencing data....59
- Overview of computational identification and DE of LncRNAs 59
- Methods....62
  - RNA-Seq and library preparation....64
  - Aligning reads to the genome....66
  - Selecting reads according to overlapping genomic features....68
  - Identify expressed regions outside known gene models....69
  - Reconstruct genes of putative LncRNAs....75
  - Calculate DE and associate expression profiles of putative LncRNAs and genes....85
  - Comparing conditions using Generalized Linear Models (GLMs) 86
  - Annotation of predicted LncRNAs....90
  - Calculate counts and DE of known genes....92
  - Functional enrichments....93

Transcriptional changes of protein coding genes and novel LncRNAs in rat's DRG after the SNT pain model.....96
Overview.....96
Background.....97
The Spinal Nerve Transection pain model.....97
RNA-Seq and library preparation.....98
Aligning RNA-seq reads to genome.....99
Experimental Design.....101
Further quality control.....101
Results.....105
Differential Expression analysis of known genes.....105
Functional enrichment.....110
Expression patterns of ion channels and pain genes.....114
Identification of LncRNAs.....120
Expression of LncRNAs in rat's DRG.....122
LncRNAs and pain-related protein coding genes.....129
Discussion.....133
Transcriptional changes of LncRNAs and protein coding genes in DRG of two mouse strains experiencing high and low induced hypersensitivity.....135
Overview.....135
Background.....136
The Spared Nerve Injury pain model.....136
Behavioural tests.....137
Mouse strains and phenotypes.....140
Dissections.....142
RNA isolation and extraction.....144
Dataset.....144
RNA-Seq and library preparation.....145
Aligning RNA-seq reads to genome.....145
Experimental Design.....146
Further quality control.....148
Results.....153
Differential Expression analysis of known genes.....153
Identification of LncRNAs.....174
Expression of predicted LncRNAs in mouse DRG.....178
Differential Expression of LncRNAs.....180
Antisense LncRNAs and pain-related protein coding genes.....187
Intergenic LncRNAs and pain genes.....189
Comparing the mouse results to rat under the SNT pain model...193
Discussion.....200
Clustering of patients with diabetic neuropathy reveals distinct neuropathic pain dimensions.....204
Overview.....204
Introduction.....204
The Neuropathic Pain Symptom Inventory (NPSI).....205

Douleur Neuropahique en 4 Questions (DN4).....208
Toronto clinical scoring system (TCSS).....210
The Quantitative Sensory Testing (QST).....212
The 7-Day pain diary.....212
Clinical markers.....213
Methods.....214
Imputing missing values.....214
Clustering.....215
Data Analysis and statistical tests.....215
Recoding variables.....217
Transform scores into categorical variables.....218
Normalization and imputation of missing values.....219
Dataset.....219
Results.....221
Distribution of pain scores across sexes and clinical markers for patients with painful neuropathy.....221
Consistency of neuropathic pain screening tests.....226
Quantitative Sensory Testing scores associated with self reported scores and clinical markers.....232
Principal Components Analysis and clustering.....241
Factor analysis and contributions to the Principal Components...245
Clustering.....254
Association of clusters and principal components to clinical markers.....260
Conclusion.....267
Conclusions and future work.....269
Identifying LncRNAs from RNA-seq data.....270
Transcriptional profiling of rodents DRG.....272
Divergent phenotypes of painful neuropathy.....274
Appendix 1.....277
Appendix 2.....278
Appendix 3.....279
Appendix 4.....281
Appendix 5.....283
Bibliography.....284

List of Figures

Introduction

1. Distinct types of nociceptors and Aβ fibers responding to light touch project on different laminae of the spinal cord……16
2. Sub-graph of significantly over-represented GO terms in pain genes…..20
3. Main rodent models of neuropathic pain……..23
4. SNI pain model……27
5. Classification of LncRNAs according to genomic context……..30
6. Overview of RNA-sequencing……44
7. Overview of the dUTP strand specific protocol……...45
8. Count modes of RNA-seq reads……...52

Methods

1. Flowchart of computational pipeline……63
2. Aligning RNA-seq reads with the STAR aligner……..67
3. Coverage graph of RNA-seq across an annotated gene model……70
4. Genomic Ranges in a GRanges object……..72
5. Islands of Expression are being grouped together and trimmed by Splicing Junctions into putative LncRNAs……...76
6. Grouping of islands of expressions and splicing junctions……..79
7. Hits of original splicing junctions (queryHits) on the disjoint splicing junctions (subjectHits)……..80
8. Example of how our pipeline identified a known LncRNA using a RNA-sequencing data……84
9. Area under the curve of a classifier built to distinguish between correlation calculated form from random pairing and correlation of actual pairs of proximal genes and LincRNAs……91

Transcriptional changes of protein coding genes and novel LncRNAs in rat's DRG after the SNT pain model

1. The SNT pain model……97
2. Ambiguous reads and multi-mappers……103
3. Boxplot of Cook's distance for ENSEMBL genes……104
4. Hierarchical clustering of samples according to regularized log2 counts of ENSEMBL genes……106
5. Principal Components analysis of regularized log2 counts of ENSEMBL genes……..107
6. Volcano plot showing the relationship between the log2 fold change of genes and the p.value……..110
7. Gene Ontology subgraph leading to the top five highly enriched GO terms……...114
8. Expression patterns of pain genes……...117
9. Expression patterns of voltage-gated potassium channels……..118
10. Expression patterns of voltage-gated sodium channels……119
11. Heatmaps of TRP, Chloride and Calcium ion channels……120

12. Distribution of exon numbers in the predicted LncRNAs……123
13. Median counts for ENSEMBL genes and predicted LncRNAs……124
14. Distribution of read counts for predicted LncRNAs…….125
15. Boxplot of Cook's distances for all predicted LncRNAs…….126
16. PCA according to the expression of predicted LncRNAs……129
17. Distances between LincRNAs and protein coding genes…….132

Transcriptional changes of LncRNAs and protein coding genes in DRG of two mouse strains experiencing high and low induced hypersensitivity

1. The SNI pain model……138
2. Mouse undergoing Von Frey filament testing for hypersensitivity in the pain lab……139
3. Von Frey filaments time course ipsilateral to the injury…….140
4. Von Frey filaments time course contralateral to the injury…….141
5. Identification of the L5 DRG…….144
6. Number of multi-mappers and ambiguous reads for each sample…….151
7. Boxplots of ENSEMBL genes' Cook's distance across all samples…...153
8. Hierarchical clustering of samples according to regularized log2 counts of ENSEMBL genes……155
9. Clustering of BALB/c (left) and B10.D2 (right) sample according to regularized log2 counts of ENSEMBL genes…….156
10. Principal components analysis for BALB/c and B10.D2 strain……...157
11. Expression pattern of pain genes based on rld transformed counts…..159
12. Common significantly DE pain genes in both strains……..161
13. Heatmaps of ion channels for BALB/c strain and B10.D2 strain……163
14. Venn diagram of significantly DE genes in mouse strains……...165
15. Log mean counts of Identified and Un-identified ENSEMBL LncRNAs in the mouse RNA-seq dataset……175
16. Distribution of exon number in the predicted LncRNAs……..176
17. Annotated LncRNAs as predicted in our dataset……..177
18. Median read counts for predicted LncRNAs compared to median read counts for ENSEMBL protein coding genes……..179
19. Clustering of samples according to the expression of predicted LncRNAs……181
20. Cook's distance of the expression of novel LncRNAs for each sample in our mouse RNA-seq dataset……...182
21. Maximum Cook's distance of LncRNAs and ENSEMBL genes…….183
22. Distribution of CPC scores for all predicted LncRNAs and for those significantly DE……..186
23. Distance between ENSEMBL genes and LincRNAs……..190
24. Area under the curve of a classifier built to distinguish between random correlation and correlation of actual pairs of adjacent LincRNAs and protein coding genes……...192
25. Oprd1 pain gene and chr4:132107492-132108725(-) adjacent LincRNA in the genome browser……193
26. Venn diagram of DE genes in B10.D2, BALB/c mice and rat…….194

Clustering of patients with diabetic neuropathy reveals distinct neuropathic pain dimensions

1. A digital version of the NPSI questionnaire as it is in the database used by the current study…….207
2. The English form of the DN4 questionnaire……..209
3. The TCSS pain questionnaire…….211
4. Females report more severe neuropathic pain than males……..222
5. No association between gender and HbA1c and IENFD……223
6. Significant correlations between NPSI scores and HbA1c and Age…..225
7. Correlations between TCSS scores and DN4 score……...228
8. TCSS sensation sub score is very highly correlated to the MRC sensory score……229
9. DN4 score is strongly correlated to NPSI average score in patients with painful neuropathy……..229
10. Correlation of TCSS scores to the NPSI average total score for patients with painful neuropathy……..230
11. The WSI and CSI did not show any significant correlation with the NPSI scores for patients with painful neuropathy……231
12. Correlation of DN4 scores to the QST parameters……...234
13. Correlation of HbA1c mmol/mol to the QST parameters…….235
14. Correlation of IENFD mmol/mol to the QST parameters……236
15. QST parameters are highly correlated to the IENFD……...238
16. QST parameters are not strongly correlated with NPSI average pain score……239
17. Principal components analysis of QST data…….242
18. Scree plots for NPSI data…….243
19. Distinct pain dimensions identified by the principal component analysis of NPSI data……244
20. Individuals factor map for NPSI data……...246
21. Correlation between age and 4 first NPSI principal components…….249
22. Females had higher values in NPSI PCs than males……250
23. Correlation between HbA1c mmol/mol and 4 first NPSI principal components…….252
24. Within groups sum of squares against the number of clusters…….254
25. NPSI average total score for each cluster…….255
26. Cluster plots colour coded by pain severity……..256
27. Hierarchical clustering of patients based on NPSI data-point……..258
28. Association between Pain Diary and DN4 scores to NPSI clusters….260
29. IENFD and HbA1c vs cluster assignment……261
30. Age and gender are highly associated with cluster assignment……263
31. TCSS symptoms sub score has a significant effect on the on grouping of patients according to clusters……..264
32. NPSI scores have significant effect on cluster assignment. Clustering revealed distinct modalities of pain obscured from the NPSI total score...265

List of tables

Methods

1. Attributes of Islands of Expression……..72
2. Log file showing all the rounds of read selection and island of expression identification…….74
3. Attributes of putative LncRNAs……...83

Transcriptional changes of protein coding genes and novel LncRNAs in rat's DRG after the SNT pain model

1. Quality controls for all 4 sequencing lanes……...99
2. The top 20 enriched GO terms……112
3. LncRNAs antisense of Kcna2 and Scn9a gene……...128
4. Significantly DE LncRNAs antisense of significantly DE pain genes...131
5. Significantly DE LincRNAs with a significantly DE pain gene as their closest neighbour……133

Transcriptional changes of LncRNAs and protein coding genes in DRG of two mouse strains experiencing high and low induced hypersensitivity

1. Quality controls for all sequencing lanes……146
2. Top 20 enriched GO terms for biological process in SNI vs Sham for BALB/c strain…….167
3. Top 20 enriched GO terms for biological process in SNI vs Sham for B10.D2 strain……..168
4. Top 20 enriched GO terms for biological process in genes with significant different response between strains……..171
5. LincRNAs with highly correlated expression to pain genes……..191
6. GO enrichment for DE genes with the same direction in both BALB/c mouse strain and rat……195
7. GO enrichment for DE genes with the same direction in both B10.D2 mouse strain and rat……196
8. Syntenically conserved LincRNAs between rat and mouse with the same pain gene as their closest gene……197
9. Conserved antisense LncRNAs on the opposite strand of Scn9a and Kcna2 genes……198
10. LncRNAs antisense of pain genes, syntenically conserved between mouse and rat……..199

Clustering of patients with diabetic neuropathy reveals distinct neuropathic pain dimensions

1. Recoding of Spontaneous Ongoing pain variable……..218
2. Recoding of Spontaneous Paroxysmal Pain variable…….218
3. Pain severity categories derived from NPSI scores……219
4. Distribution of NPSI scores across genders……221

Introduction

This thesis presents a number of studies in the general subject of Bioinformatics and Functional Genomics. The studies presented were made in collaboration with experimental scientists of the London Pain Consortium (LPC), an initiative that has promoted collaborations between experimental and computational scientists to to further our understanding of pain.

The work presented in this thesis is mainly concerned with the molecular biology of pain and deals with data gathered from high throughput technologies aiming to assess the transcriptional changes involved in well induced pain states, both from animal models of pain and human patients. The ultimate aim is to gather valuable insights that will help us understand pain and more specifically neuropathic pain and produce efficient drugs to control it. In order to do so we have used RNA-sequencing data of tissues involved in the nervous system, clinical markers and self-completed pain questionnaires.

Profiling technologies have been used extensively in the context of pain research to identify Differentially Expressed (DE) genes under well-induced pain states, usually using animal models. In addition advancements in next generation sequencing and particularly in RNA-sequencing (RNA-seq) have enabled us to comprehensively assess transcriptional changes of animal models under specific pain models. Thus, we are now able to detect alterations in expression for both annotated and un-annotated genomic regions and to identify Long Non-Coding RNAs (LncRNAs), which may contribute to neuropathic pain.

In this thesis we present a study the primary aim of which is to identify pain-related genes and transcriptional patterns of pain, novel pain-related genes encoding for LncRNAs, which may contribute to neuropathic pain and to analyse their biological pathways using functional genomic approaches. Moreover we will analyse both molecular and clinical data in

order to gather more insights regarding the different qualities of neuropathic pain and to further understanding of its molecular signature.

Pain

Pain has played a crucial role in human evolution as its sensation protects the body from serious injury. The ability to detect and respond to such stimuli is crucial for surviving (Basbaum et al., 2009). In other words the sensation of pain gives to the organism an early warning about a potentially damaging stimulus (Woolf and Salter, 2000). In 1968 in his paper “Psychological aspects of pain” (Merskey, 1968) H. Merskey gave the following broad definition of pain: “an operational definition of pain should be adopted as follows: ‘An unpleasant experience which we primarily associate with tissue damage or described in terms of such damage, or both’”. As the International Association for the Study of Pain rephrases it: "Pain is an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage."

Pain it is not homogenous however, it can be due to an external stimuli or due to a malfunction of the nervous system. In general pain involves higher level emotional components and it qualitatively differs from the body’s response to a potentially damaging stimuli. IASP classifies pain according to its features or more specifically by the region of the body involved, the system whose dysfunction may be causing the pain, the duration and pattern of occurrence, the intensity and the aetiology. But other scientists argue that pain can be classified into three main categories: physiological, inflammatory and neuropathic (Woolf and Salter, 2000). Physiological pain is nociceptive pain, it is activated by a noxious stimuli. Inflammatory pain is pain due to the response of the immune system and inflammation and neuropathic pain is pain due to damage or disease affecting the nervous system. Both the peripheral and the central nervous system are involved in pain. Pain can be acute, due to an intense stimuli when detection coding and modulation of noxious stimuli generates pain or persistent – chronic pain. Persistent pain involves increased plasticity of the

pain transmission pathway leading to hypersensitivity, i.e. the peripheral and/or central nervous system enhancing pain signals. Neuropathic pain is indeed the verbalization of such maladaptive neuronal plasticity, in the context of trauma or lesions to the somatosensory nervous system (Costigan et al., 2009). As mentioned above pain can be initiated, and physiological pain generally is, by a noxious external stimuli. The process of detecting encoding and processing such a stimuli is nociception. This process is carried out by the nervous system and it involves the encoding of a broad range of mechanical, chemical and thermal stimuli. The nerve fibres which detect such stimuli are called nociceptors and lie in the periphery of the nervous system (Basbaum et al., 2009). Nociceptors transmit information to neurons in the spinal cord which in turn transmit information to the cortex via their projections and create the sensation of pain (Gold and Gebhart, 2010). Thus nociception is the detection of thermal, mechanical and chemical stimuli from high threshold nerves on the peripheral nervous system, on the other hand the sensation of pain *per se* involves higher functions of the brain in order to process this information.

The cell bodies of nociceptors innervating the body are located in the Dorsal Root Ganglia (DRG), and those of nociceptors innervating the face are in the trigeminal ganglion. In general, axons of nociceptors are connecting target organs to the spinal cord. Contrary to the typical neuron, nociceptors can facilitate bidirectional transmission of information, as proteins expressed in DRG or trigeminal ganglion are transferred from the dendrite to the axon and vice versa.

Nociceptors are classified into two major classes: medium diameter myelinated (Aδ) afferents that mediate highly-localised, acute, fast pain and small diameter un-myelinated “C” fibres that mediate diffused and slow pain. More specifically these classes are divided in subclasses. For Aδ fibres type I do respond to mechanical and chemical stimuli but are not that sensitive to heat stimuli and type II have low threshold in heat stimuli but are not sensitive to mechanical stimuli. Most C fibres can detect mechanical and heat stimuli, while other are sensitive to heat and have very high

mechanical thresholds and only lower their mechanical threshold as a response to inflammation when injured. These different and specialised types project on different laminae of the dorsal horn of the spinal cord. Aδ fibres innervate laminae I and V, where laminae V receives both noxious and innocuous stimuli. C fibres innervate laminae I and II.

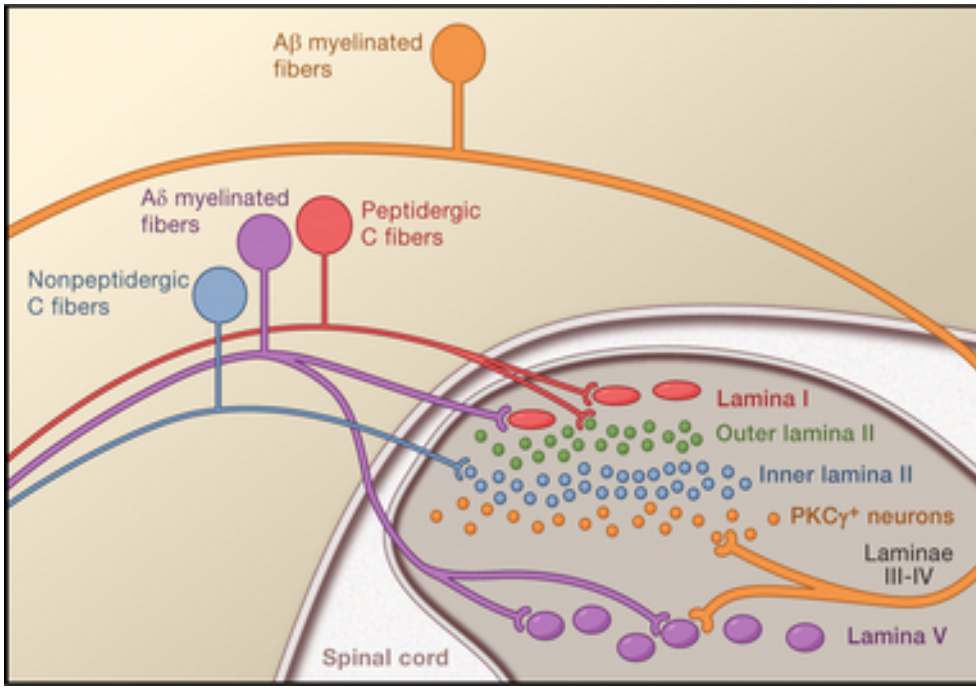


Figure 1: Distinct types of nociceptors and Aβ fibers responding to light touch project on different laminae of the spinal cord. Image courtesy of (Basbaum et al., 2009)

Pain at the molecular level

At the molecular level, pain is mediated by proteins that translate signal detection to electrical current in order to communicate with the central nervous system. Various molecules are involved in signal detection and transduction. The specialisation of nociceptors is mirrored in the molecular level. C fibres can be peptidergic, producing neuropeptides, substance P, Calcitonin Gene Related Peptide (CGRP) and expressing neurotrophic tyrosine kinase TrkA, or non-peptidergic (Basbaum et al., 2009). Non-peptidergic c fibres express the transmembrane receptor tyrosine kinase C-ret (Rearranged during Transfection) neurotrophin, G protein coupled receptors, neurterin and artemin and purinergic receptor

subtypes. Certain families of proteins are also differentially expressed between these specialised nociceptors and mediate distinct qualities of pain, ASICs respond to acidic environment, TRPV1 channels respond to heat, TRPM8 to cold and TRPA1 to chemical stimuli.

Regarding detection of noxious mechanical stimuli Transient Receptor Potential (TRP) channels are thought to activate nociceptors. TRPA1 functions as a detector of mechanical stimuli, TRPV4 is involved in pain hypersensitivity after injury and TRPV2 can respond to noxious heat and mechanical stimuli. Also KCNK potassium channels, like KCNK2 and KCNK4 and also KCNK18 act as regulators of the duration and excitability of action potentials. ASIC, acid sensitive channels, are also thought to be implicated in detecting mechanical stimuli (Basbaum et al., 2009). In addition mechanically activated ion channels are thought to be critical in initiating touch sensation and transducing mechanical stimuli. Piezo2, which is a mechanically activated ion channel expressed Merkel cells of the dorsal root ganglion has been found to be the major transducer for touch sensation (Ranade et al., 2014). Moreover, not yet identified mechanically activated ion channels, are likely to transduce noxious mechanical stimuli (Ranade et al., 2014).

After nociceptors detect pain the signal need to be transduced to the central nervous system. Voltage gated ion channels are essential for this process. Sodium and potassium channels generate action potentials in order to transmit signals to dorsal horn. Sodium channels transmit information from the periphery to the dorsal horn and potassium channels act as breaks on excitability. Calcium channels release neurotransmitters and play a key role in transmitting neuronal sensation either to generate pain or neurogenic inflammation, thus they are highly relevant to neuropathic pain. These voltage gated ion channels are crucial in modulating the excitability of nociceptors and transmitting pain from the peripheral nervous system.

Several studies have highlighted the importance of voltage gated sodium and potassium channels in pain. The Nav1.7 voltage gated sodium

channel (Koenig et al., 2015), encoded by the Scn9a gene, and Kcna2 (Zhao et al., 2013) gene that encodes a voltage gated potassium channel, are both known to be implicated in pain and also have known antisense LncRNAs regulating their expression. Recently we have published a study regarding the role of Nav1.7 – Scn9a, where loss-of-function mutations in that gene cause congenital insensitivity to pain in humans and mice (Minett et al., 2015). On the other hand, gain of function mutations of the Scn9a gene lead to hyper-excitability and intense burning sensations related to the erythromelalgia, paroxysmal pain disorder syndromes and small fibre neuropathy (Bennett and Woods, 2014; Estacion et al., 2008; Fertleman et al., 2006).

Regarding persistent pain, a group of cells and signalling molecules driving peripheral sensitisation, like non-neuronal cells related to inflammation which are recruited and infiltrate areas of tissue damage, play a crucial role. This is often called inflammatory soup (Calvo et al., 2012) and acts as mediator of peripheral sensitisation. Activated nociceptors and those non-neuronal immune cells express various signalling molecules like substance p, CGRP, bradykinin, neurotrophins, cytokines and chemokines (White and Wilson, 2008). Certain nociceptors, are activated from these endogenous chemokines and cytokines and express TRPA1, TRPV1 and ASIC channels. In addition NGF, a well known nerve growth factor implicated in embryonic development of neurons, is also an important endogenous factor of this inflammatory soup. It is expressed after nerve injury and acts on peptidergic C fibres mediating mechanical hypersensitivity after nerve injury.

Moreover, a set of manually curated genes that are validated in transgenic knockout mice to be involved in pain are available in the Pain Genes database (Lacroix-Fralish et al., 2007). This set of 430 genes represents a comprehensive repertoire of the significant transcriptional changes involved in pain. Genes are included if they found to be statistically significant differentially expressed between mutant mice (showing increased injury or stimulus induced hypersensitivity or stress or drug induced

inhibition of nociception) and wild type controls. Functional genomics examination of these genes in the context of the biological process they are involved in produced the “wheel of pain”, showing over-represented Gene Ontologies terms that describe biological processes of pain

(Lötsch et al., 2013).

Gene Ontology (GO) is a hierarchy of terms, represented as a direct acyclic graph, where each node is a clearly defined term related to biological process, cellular component or molecular function. Genes are assigned to GO terms by manual literature curation. In terms of biological processes pain is found to be associated at the molecular level to terms stemming from response to external stimulus. The 12 more highly associated terms of the genes validated to be implicated in pain (pain genes) are behaviour, response to wounding, response to organic substance, cellular ion homeostasis, ion transport, synaptic transmission, G-protein coupled receptor protein signalling pathway, intracellular signal transduction, positive regulation of biological process, regulation of system process (multicellular), anatomical structure development, regulation of localization (figure 2). Moreover cognition emerges as an over-represented process, connected with memory via learning, representing the higher brain function component of pain.

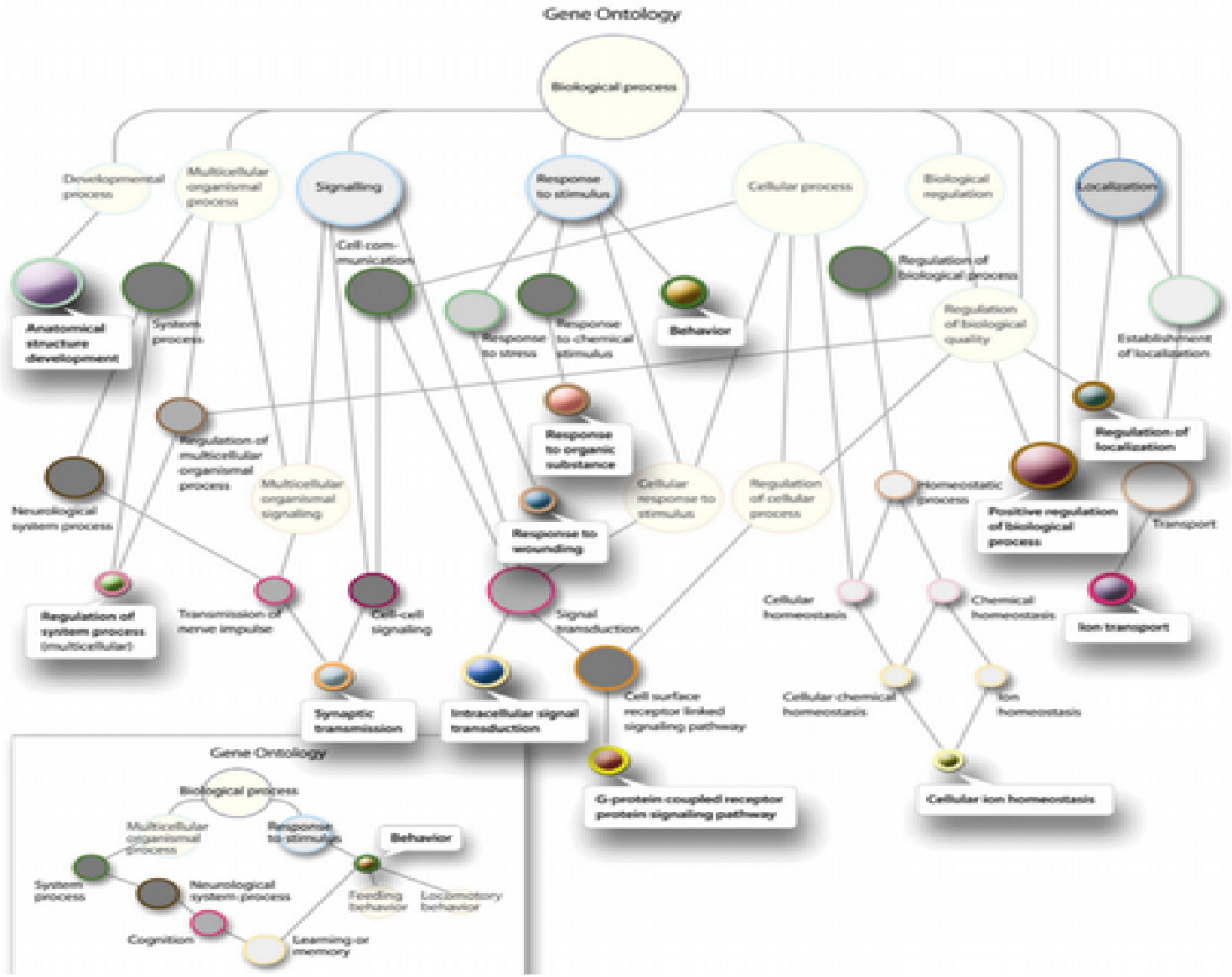


Figure 2: Sub-graph of significantly over-represented GO terms in pain genes. Image courtesy of

(Lötsch et al., 2013)

Neuropathic Pain

As the International Association for the Study of Pain (IASP) defines it, Neuropathic Pain is “pain initiated or caused by a primary lesion or dysfunction in the nervous system.” (Treede et al., 2008). Thus the diagnosis of neuropathic pain necessitates the identification of an underlying disease or lesion on the somatosensory system, which in turn gives rise to neuropathic pain as a symptom. Although neuropathic pain affects more than 5% of the population there is no adequate treatment available. Neuropathic pain causes severe dysfunctions and disabilities to patients. These come as an effect of various sensory abnormalities associated with neuropathic pain. These sensory abnormalities include persistent pain and paraesthesia, allodynia, hyperalgesia and loss of sensation (Calvo et al., 2012). Paraesthesia is the sensation of tickling, pricking or burning pain like “pins and needles” with no apparent cause, allodynia is pain evoked by a normally non-painful stimuli like brushing, hyperalgesia is the abnormally increased sensitivity to normally painful stimuli (Calvo et al., 2012).

Neuropathic pain is a symptom of an underlying neuropathy; a neuronal injury that is usually associated with “a trauma, infection, toxins or metabolic agents” (Calvo et al., 2012). In all of the above cases the organism’s response to the injury involves a robust immune response. Under pain conditions, immune cells, macrophages and monocytes are recruited to participate in the inflammatory response. These immune cells infiltrate the nerve itself, therefore, neuropathic pain has an inflammatory component, which contributes to the maintenance of pain.

Tools for assessing neuropathic pain

In order for the optimal treatment to be delivered, neuropathic pain must be correctly diagnosed. This can be a difficult task as this kind of pain is due to highly heterogeneous clinical conditions. Due to this heterogeneity of causes and clinical symptoms there is an emerging need to classify patients of neuropathic pain accordingly and provide them with the most effective treatment. For the accurate diagnosis of neuropathic pain numerous subjective pain questionnaires have been developed. As the developers of

The Neuropathic Pain Symptom Inventory (NPSI) questionnaire state “In this context, we thought it would be of interest to develop and validate a specific self-completed questionnaire for the assessment of the different symptoms of neuropathic pain. Ideally, such a questionnaire could represent a useful and exploitable tool for large cohorts of patients in multicentre studies and give information comparable to that provided by quantitative evaluation, as regards the nature and intensity of the various painful symptoms. ” (Bouhassira et al., 2004) .

Moreover a compendium of clinical assays has been systematically used to asses neuropathic pain and/or identify the underlying disease or lesion causing it. Standard neurophysiological assays can identify and quantify neuropathy. Laser evoked potentials, which specifically stimulate pain afferents, are used as diagnostic tools for central and peripheral neuropathic pain. Also punch skin biopsy can quantify the extent of skin innervation and more specifically Aδ and C nerve fibres by measuring the density of intra-epidermal nerve fibres (IENFD) (Cruccu and Truini, 2009).

Other clinical tests include the Quantitative Sensory Testing (QST) a test designed to measure response to controlled sensory stimuli. Usually Von Frey filaments, i.e. nylon filaments that will buckle elastically at a specific force measured in grams, vibrameters, i.e. plates vibrating in specific frequencies, weighted needles and thermodes, i.e. a probe that can heat or cool skin, are used in order to quantitatively assess thermal and mechanical allodynia and hyperalgesia.

Although these tools can effectively assess pain, they cannot always distinguish nociceptive from neuropathic pain. For that reason, several self reported pain questionnaires and calendars have been introduced and used extensively in pain studies aiming to screening neuropathic pain and assess its intensity. The most well known of them which have been widely used by medical doctors and researchers all over the world are DN4 (Bouhassira et al., 2005), TCSS (Bril and Perkins, 2002), the 7-Day pain diary and NPSI (Bouhassira et al., 2004).

Animal models of pain

In order to study neuropathic pain, to define its causes and develop drugs, several animal models have been developed. Pain involves higher brain functions and is subjective in both humans and other animals and as we discuss below the measurement and assessment of pain remain a challenge (Mogil et al., 2010). But at the molecular level, the interrogation of certain molecular changes involved in nociception or in processes related to chronic pain are well conserved (Khuong and Neely, 2013) not only between mammals, for example mutations in TrkA causing congenital insensitivity to pain in human were first discovered in Ntrk1-knockout mouse (Mogil, 2009), but also between distant species like drosophila, mouse and human (Neely et al., 2010). We should note that animal models of pain are exactly models of pain, they model conditions and create pain phenotypes similar but not identical to pain experienced due to diverse origins. Thus when using these models we should be very careful in translating molecular findings into clinical drugs. Regarding neuropathic

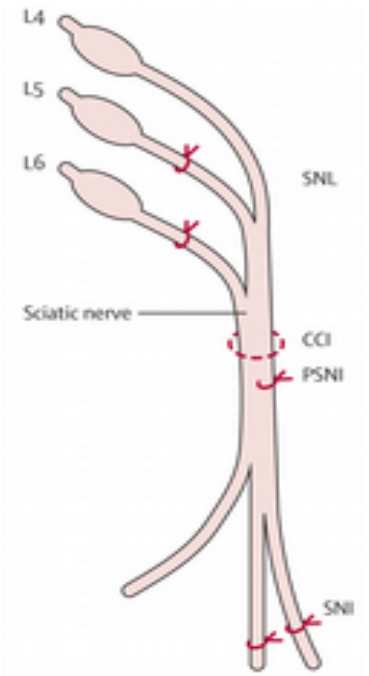


Figure 3: Main rodent models of neuropathic pain. SNL involves the ligation of L5 and L6 spinal nerves, CCI involves a loose ligation of the sciatic nerve, PSNI involves the ligation of about half of the sciatic nerve and SNI involves the ligation of the tibial and peroneal branches while leaving the sural nerve intact or other combination of two ligated and one intact branch. Image courtesy of (Calvo et al., 2012b)

pain, animal models can produce neuropathic pain-like hypersensitivity or other neuropathic pain-like behavioural.

Most of them involve rodents, usually mice and rats, and emulate neuropathy by inducing nerve injury usually through a surgical process, although chemical lesion models are also used. These models aim to produce a reproducible pain-behaviour that could be quantified by behavioural tests. The well-induced pain states measured to quantify neuropathic pain are heat-hyperalgesia, mechano-hyperalgesia, mechano-allodynia and cold-allodynia (Bennett et al., 2003). The most common animal models of pain include the Chronic Constriction Injury (CCI), the Partial Sciatic Ligation (PSL), the Spinal Nerve Ligation (SNL), Spared Nerve Injury (SNI) and the Spinal Nerve Transection (SNT). In our study we will mainly use data derived from mice and rats that have undergone the SNT and SNI pain models and humans with diabetic neuropathy which is related to neuropathy and neuropathic pain. An overview of the main models of neuropathic pain evoked by peripheral nerve injury can be seen in figure 3.

The usage of animal models has been pivotal for our further understanding of pain. As described above by inducing sensory abnormalities in animals several human conditions leading to pain and neuropathic pain can be modelled. Thus a series of physiopathological phenomena can be modelled by a diverse array of animal models, which have been developed to model the diverse aetiology and manifestations of neuropathic pain observed in humans. Moreover different models and consequently different behavioural assays model different types of pain, namely acute nociceptive pain, spontaneous pain, hypersensitivity and allodynia due to peripheral or central nerve injury, drug induces or disease induced pain (Jaggi et al., 2011).

As described above there is an ensemble of neurophysiological tools, clinical protocols and self-completed questionnaires that can be used in a research or clinical context in order to assess and quantify pain. Nonetheless

these do not answer any reductionist question about possible causes of pain in the molecular level, a question very relevant for scientific research, or they do not identify any objective biomarker that could be used in order to identify pain in the molecular level (Mogil, 2009).

Animal models of pain can overcome limitations of experiments on humans as research subjects in the field of pain. Usage of animals, which have undergone some invasive or pharmacological process in order to facilitate a well induced pain state has been essential for pain research. Additionally well replicated controlled experiments, stratified for several factors are much more feasible when using non-human animal models of pain. Moreover we can harvest tissue and carry out molecular assays in order to understand transcriptional or translational changes in pain states, where usually harvesting pain-relevant tissue involves dissection.

Several tests have been introduced to measure for different dimensions of pain. Most of those methods involve the application of a noxious stimuli to a body part of the animal, usually a hind-limp, and then the recording of a simple pain related behaviour that can be scored, usually limp withdrawals. Of course behavioural tests can be much more complex in order to score persistent pain, those tests involve long term recording of normal behaviours like mating, looking for food, guarding, biting etc. Most of the tests measure spinal reflexes or simple natural behaviours and are not always relevant to pain *per se* (Mogil, 2009). Additionally, most of the tests are biased towards the measure of hypersensitivity, i.e. allodynia and hyperalgesia, because they record and score evoked responses when in reality many patients with chronic pain suffer paroxysmal pain. As a matter of fact, in this study we used Von Frey filaments which measure induced hypersensitivity or allodynia after a pain model of peripheral neuropathy.

We should state here that in animal models of pain we cannot directly measure pain. Instead we can measure several behavioural aspects indicating pain. On the other hand this is also the case for humans, as we also rely on behavioural tests or questionnaires exploiting the advantage of

linguistic communication between humans. J. Mogil has discussed these aspects of animal models of pain, presenting the advantages and drawbacks, but generally highlighting the necessity of animal models in the research of pain (Mogil, 2009; Mogil et al., 2010). In the current study we used rodent models of pain, namely rats and mice.

The Spinal Nerve Transection (SNT) Neuropathic Pain Model

As mentioned above one of the main models for neuropathic pain is the Spinal Nerve Transection model. According to that specific model protocol a well-induced pain behaviour is imposed on the animal by a surgical injury to the Lumbar 5 or 6 of the Dorsal Root Ganglion. The injury consists of a tight ligation and transection of the L5 or L6 spinal nerve (Bridges et al., 2001), figure 3. Thus the SNT model is a model of peripheral neuropathy which leads to long-lasting mechanical and thermal hypersensitivity as well as to hyperalgesia (Bennett et al., 2003). As a control group it is possible to use animals which undergo only sham surgery. In this case the spinal nerve is being exposed but not ligated.

The Spared Nerve Injury (SNI) pain model

SNI is another well known model of pain that has been proposed in order to achieve a well induced and reproducible neuropathic pain phenotype. The goal is to produce a pain model which results in reproducible sensory abnormalities including allodynia, hyperalgesia and spontaneous bursts of pain (Jaggi et al., 2011) in a period of time broad enough to allow extensive behavioural, clinical and molecular assays to be implemented. SNI is a relatively modern model of neuropathic pain, proposed in 2000 by Decosterd and Woolf (Decosterd and Woolf, 2000). In this model two of the sciatic nerves are axotomised and one is left untouched, i.e. spared, thus the model's name Spared Nerve Injury. In order for the model to be implemented correctly huge caution is needed in order to not injure the untouched nerve. Different combinations exist, tibial and common peroneal axotomised and sural spared or the other way around. Sparing the tibial branch (figure 4) produces consistent and robust

mechanical allodynia without increasing heat sensibility (Shields et al., 2003).

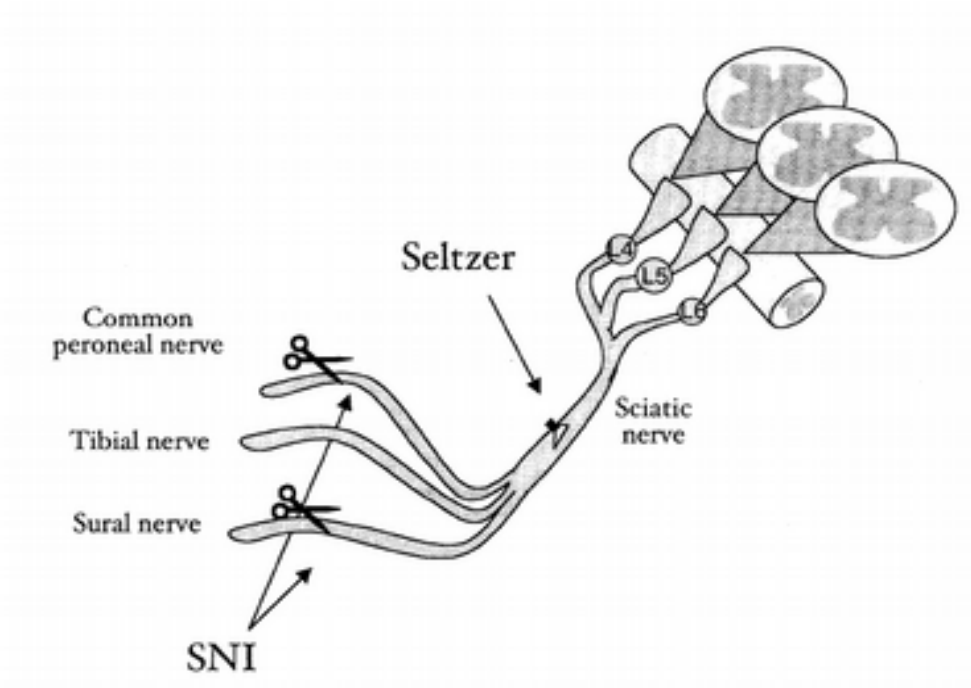


Figure 4: SNI pain model, the common peroneal and sural nerve branches are axotomised. The tibial nerve is spared.

Gene Expression

The Central Dogma

As Francis Crick states in his seminal paper "Central Dogma of Molecular Biology" (Crick, 1970) the flow of genetic information could be formulated as follows: DNA is transcribed to RNA which is consequently translated to Proteins. Similarly, in F. Crick's words, "*The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid.*".

The central dogma of molecular biology describes the flow of genetic information within a biological system as follows: DNA replicates itself, DNA is transcribed to RNA, then RNA is translated to protein. The unit of heredity that codes for a protein is defined as a gene. Yet a very small percentage of a genome actually codes for proteins. In this study we will also study sequences which are transcribed but not translated into polypeptides. More specifically we will deal with a subclass of non-coding RNA. The vertebrate's genome is transcribed into a large repertoire of non-coding RNAs. Most of them are small and this class of small RNA includes the small nuclear/ nucleolar RNA (sn/snoRNA), the miRNA, siRNA and piRNA all of which generally regulate expression activity. There are also the classes of ribosomal RNA (rRNA) and transfer RNA (tRNA). The class which we will study more in depth is that of long non-coding RNAs, these are non-coding RNAs which are more than 200bp long.

Long Non-coding RNAs (LncRNAs)

Although LncRNAs is a topic that attracts a lot of focus in current research there is not a general and unambiguous way of defining and identifying them (Kapusta and Feschotte, 2014). Recent papers have proposed bioinformatics strategies to identify transcripts of LncRNAs but they use slightly different definitions and properties for them. One of the main properties of LncRNAs is that they are mainly defined by a set of negative traits (Ulitsky and Bartel, 2013) like non-coding for proteins and not belonging in other families of small non protein coding RNAs.

Long non-coding RNAs can be intergenic, antisense i.e. on the opposite strand of protein coding gene models overlapping any exon, sense overlapping i.e. overlapping introns and exons of a protein coding gene and producing a non-coding transcript, intronic i.e. completely nested within an intron of a protein coding gene or divergent i.e. when its transcription is initiated by a bidirectional promoter common with a protein coding gene in very close genomic proximity.

In the rest of our study we refer to Long Non-Coding RNAs as LncRNAs, and we specifically focus on intergenic ones, namely between known protein coding gene models, and antisense ones, namely on the opposite strand of known gene models. Thus we classified LncRNAs according to their genomic context while acknowledging that LncRNAs could also be classified by non-mutually exclusive criteria according to their chromatin content, subcellular localization, structures and function. We should also note that different studies propose slightly different names, which describe the same classes of LncRNAs (Harrow et al., 2012; Ponting et al., 2009; Ulitsky and Bartel, 2013, 2013). Using the genomic context classes, presented in the comprehensive review papers “Volatile evolution of long noncoding RNA repertoires, mechanisms and biological implications”, “Evolution and Functions of Long Noncoding RNAs”, “lincRNAs: Genomics, Evolution and Mechanisms” we focus on what are consistently described as intergenic and antisense LncRNAs (Figure 5).

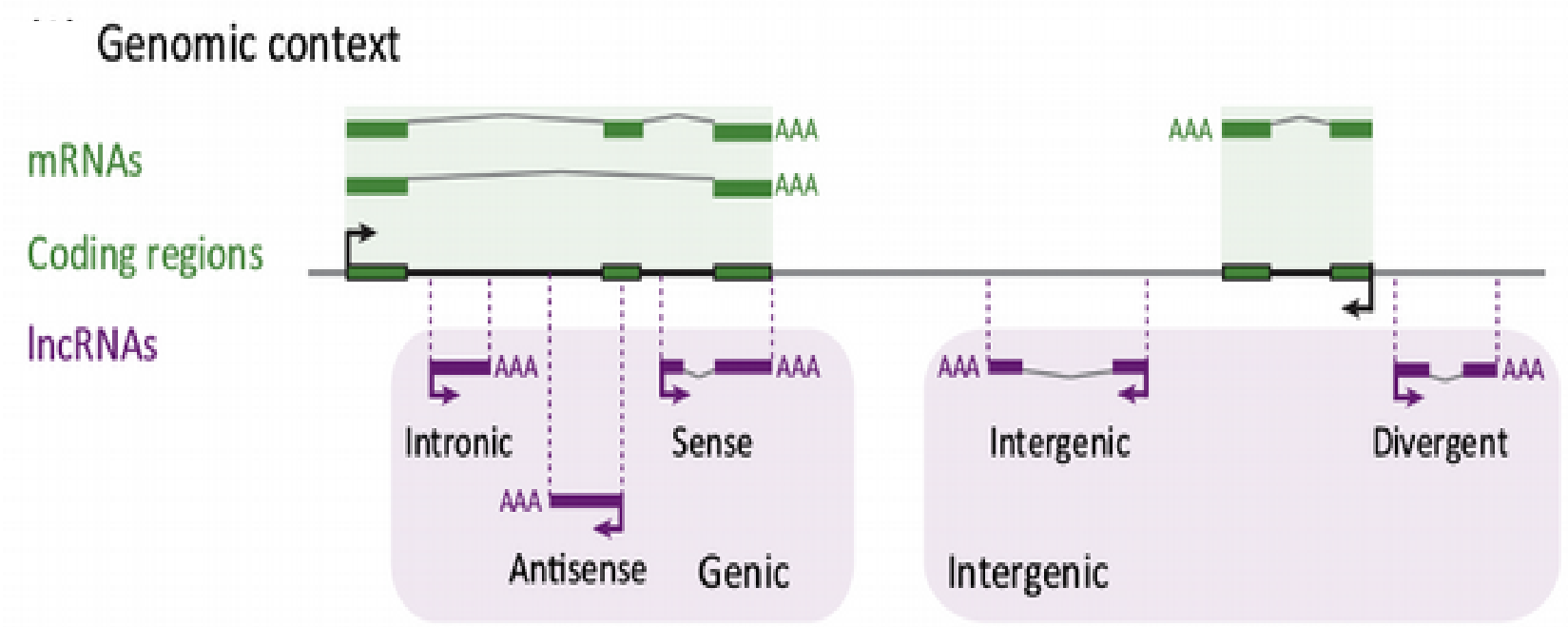


Figure 5: Classification of LncRNAs according to genomic context. (Kapusta and Feschotte, 2014)

The main characteristic of the LncRNAs is that they are usually 5' capped, non-protein coding, polyadenylated transcripts of more than 200bp long. They are usually multi-exonic even though they do not code for proteins. Various studies, including the GENCODE annotation comprising of manually curated and predicted gene models, show that they usually have two exons (Harrow et al., 2012; Ilott and Ponting, 2013; Marques and Ponting, 2014; Ponting et al., 2009; Young et al., 2012), which are slightly longer than those of protein coding genes. In mammalian genomes several studies have predicted vast quantities of LncRNAs and pervasive transcription, 1% of human's genome codes for the exonic parts of LncRNAs alone, although different methods and criteria produce highly divergent catalogues of LncRNAs. Some catalogues are thought to have a significant percentage of putative LncRNAs produced by computational or experimental artefacts, namely pseudogenes, mRNA precursors, sequencing noise in the form of scattered mono-exonic transcripts or genomic contamination (Ponting et al., 2009).

Less than 10% of the human genome encodes for mRNAs and spliced non-coding RNAs, out of this only 1% encodes for proteins, leaving about 9% transcribed into non-coding transcripts but with yet unknown function (Ponting et al., 2009), it is also thought that at some point every single genomic base of most mammalian organisms would be transcribed (Marques and Ponting, 2014). In addition about 30% of protein coding genes have a natural antisense transcript in the opposite DNA strand. It is such the extent of non-exonic, non-canonical transcription that in our dataset only about 20% if RNA-seq reads were mapped to annotated exons.

LncRNAs are thought to have different roles in regulating gene expression (Marques and Ponting, 2014) although due to the general lack of conservation most of them are thought to be non-functional (Ulitsky and Bartel, 2013). The functional ones can regulate expression in cis, i.e. regulating neighbouring genes, or in trans, i.e. regulating distal genes. In addition some of the non-functional ones may regulate gene expression by just being transcribed in the first place, where the transcription itself can

interfere with chromatin changes or with the transcriptional machinery itself which in turn regulate gene expression. In that case the transcription itself regulates expression and not its product (Ulitsky and Bartel, 2013).

In terms of sequence conservation they are under modest negative selection with a very low average selective constraint. On the other hand, although they are not usually conserved across species, they show syntenic conservation as they can be found in genomically equivalent positions across different species. Usually less than 5% of their sequence is conserved (Marques and Ponting, 2014) and that leads to the hypothesis that even a very small fraction of sequence can be enough to retain function for the functional ones.

In terms of expression levels they are notoriously lowly expressed, they can have a median expression 10 fold lower than that of protein coding genes (Ilott and Ponting, 2013; Marques and Ponting, 2014; Ponting et al., 2009; Ulitsky and Bartel, 2013). In addition they have very diverged expression patterns, their expression is highly cell type and developmental stage specific, thus it is hypothesized that they contribute to the diverged pattern of transcriptional changes emerged (Marques and Ponting, 2014).

Given all these features, it is inherently difficult to identify novel LncRNAs. Very indicative, different studies have described sets of LncRNAs with very low overlap. For example GENCODE's (Harrow et al., 2012) human annotation, which has the most annotated LncRNAs, shows very modest overlap with other studies. Namely 42% of the GENCODE LncRNAs intersect with the database lncRNAdb (Amaral et al., 2011) and 39% with the catalogue published by Cabili (Cabili et al., 2011).

Young et al (Young et al., 2012) predicted 119 LncRNAs in drosophila while the modEncode project (Brown et al., 2014) collected evidence for 3504 transcribed regions. From those datasets only 246 LncRNAs gene models could be constructed. This is a result of several technical difficulties such as the very low expression of the transcripts, lack of data for the Transcription Start Site (TSS), poor genome annotation, high

tissue specificity of the transcript etc. In human GENCODE v13 (Harrow et al., 2012), the reference human annotation from the ENCODE/ENSEMBL team, predicted 12393 LncRNAs while Cabili et al (Cabili et al., 2011) predicted 8263. Only 4343 LncRNAs are common in these sets, this lack of agreement again arises from some of the barriers presented above.

In addition there is limited overlap and significant differences in gene models annotated by ENSEMBL/GENCODE (Yates et al., 2016) and RefSeq (Pruitt et al., 2014) consortia. For these reasons there is group of studies which have focused only in LncRNAs showing some splicing activity in order to control for RNA-sequencing noise or genomic contamination, or in certain sub classes of LncRNAs, i.e. only lincRNAs (Marques and Ponting, 2014; Ulitsky and Bartel, 2013). In addition, as there are no sequence or genomic characteristics that can be directly used to infer function, a divergent expression profile and transcriptional changes under certain conditions, cell types or developmental stages are good indicators of functional LncRNAs (Ponting et al., 2009) .

The review of Ponting and Illott (Ilott and Ponting, 2013) presents a comprehensive analysis of the main methods for predicting long non-coding RNAs using RNA sequencing. The classic workflow for predicting LncRNAs is described as follows: After RNA extraction and library preparation comes sequencing by synthesis. During sequencing, strand information can be either retained or not, while reads can be paired-end or single-end. Then reads are aligned to a reference genome. In order to remove protein coding transcripts all the known, annotated genomic regions which are included in one or more reference gene sets are excluded. This is followed by the identification and assignment of reads to what we call islands of expression, namely genomic regions outside gene models which accumulate reads over a certain threshold, or Transcriptionally Active Regions (TARs) (Gerstein et al., 2014). We should note here that these islands of expression could be either non-coding RNAs, yet unknown protein coding genes, erroneously mapped reads to the genome, sequencing artefacts or genomic contamination. Thus every pipeline for identifying

LncRNAs from RNA-sequencing uses a lot of filtering downstream of the identification of TARs or islands of expression. The next step involves selecting transcripts by their size, usually more than 200bp length. As we do not have annotated gene models it is only possible to identify Islands of Expression that are likely to represent novel exons, or groups of novel exons in close proximity so as to retain a certain coverage threshold. Subsequently the sequences of the above transcript set are fed into algorithms that assess their coding potential.

Another way to reconstruct gene models is to perform a complete transcriptome assembly and then to discard transcripts derived from known protein coding gene models. There are tools like Cufflinks (Trapnell et al., 2012), Scripture (Guttman et al., 2010) and StringTie (Pertea et al., 2015) which use a splice graph to identify transcripts and can reconstruct the whole transcriptome relying on RNA-sequencing data and a genome assembly (Ilott and Ponting, 2013). StringTie was released in 2015, after the completion of most of the study presented in this thesis and despite the existence of the two other methods, most of the studies aiming to the identification of LncRNAs have developed some computational pipeline tailor made to the exact task of the identification of LncRNAs (Xu et al., 2016). Although Cufflinks has been used in a lot of studies aiming to identify novel isoforms, it does not produce an annotation which is suitable for count based RNA-seq analysis at the gene level. As a matter of fact, its transcriptome reconstruction step produces a set of transcripts with FPKM scores assigned to them and calculates DE based on those scores. Thus its output can only be analysed by Cuffdiff (Trapnell et al., 2012).

In general, these tools focus on extending the known annotations, so they can identify un-annotated isoforms. Then, they produce a set of full-length transcripts, but they are computationally expensive as they do not focus only on identifying novel LncRNAs. All these are not drawbacks of those methods, but rather direct consequences stemming from the original aim of reconstructing the whole transcriptome and not just identifying novel LncRNAs. There are also tools that can *de-novo* assemble RNA-seq reads to

perform a whole transcriptome reconstruction, but these *de-novo* assemblers are not well suited for identifying LncRNAs as they are very lowly expressed (Ilott and Ponting, 2013).

Recently numerous studies reported pervasive transcription to be prominent at similar levels in most species (Gerstein et al., 2014). As an example 32% in Human, 36.9% in Worm and 34.5% in fly sequenced genomes were transcribed into transcriptionally active regions of no-canonical transcription excluding all classes of annotated non-coding RNAs, mRNAs and pseudogenes. The authors call the Islands of Expression where non-canonical, non-protein coding, transcription takes place Transcriptionally Active Regions (TARs). That is an accurate description of the regions able to be identified using RNA-seq for analysing non annotated parts of a genome.

In our study we used the term Islands of Expression for these regions and we found that 30% of our RNA-seq reads were mapped to yet unknown parts of the genome. Regarding annotated genes, 21857 ENSEMBL genes were expressed in mouse DRG, associated with 86525 transcripts, out of which 42685 (49.33%) were non-coding transcripts of various classes on non-coding RNAs.

Several thorough studies aiming at the identification of LncRNAs give us insights about the details of the relevant bioinformatics pipelines and produce an indirect but comprehensive set of properties which define LncRNAs. For example in studies including the “Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses”(Cabili et al., 2011) and “Identification and Properties of 1,119 Candidate LincRNA loci in Drosophila melanogaster genome” (Young et al., 2012), the main defining feature of LincRNAs is that they are long transcripts of more than 200bp and 1kb on average, with very small or no coding potential. They are on average shorter than coding gene models, simpler and with fewer exons, but most of them are multi-exonic. They are expressed in lower levels than protein-coding genes. Additionally studies

like the “Comparative analysis of the transcriptome across distant species” (Gerstein et al., 2014) present a comprehensive pipeline for identifying transcriptionally active regions and predicting LncRNAs from chromatin signatures. Another approach is to use supervised predictions i.e. to build classifiers which integrate all those different features to distinguish actual LncRNAs from false positives. However these approaches can only detect about 10% of actually present transcripts in Human (Gerstein et al., 2014). These results suggests that whole new classes of LncRNAs might exist. Moreover studies like the “Diversity and dynamics of the Drosophila transcriptome” (Brown et al., 2014) present us with methods for reconstructing LncRNAs gene models.

Although, as discussed above, databases and annotated sets of LncRNAs have been disseminated for various species, like Human, Mouse, Fly and Worm, those classes of annotated LncRNAs may dramatically differ from the novel islands of expression, or TARs, identified. This finding suggests that there might be a huge repertoire of functions and expression profiles in non-coding RNAs yet to be discovered.

Functional repertoires of LncRNAs

As we previously discussed certain features of LncRNAs suggest that most of them are non-functional. Also it is hypothesized that cells transcribe regularly and then control expression mostly post-transcriptionally. From an evolutionary perspective there is no point of spending energy in order to stop transcription of DNA as protein abundance or abundance of non-coding RNAs can be post-transcriptionally regulated or non-functional transcripts can degrade rapidly without any consequence (Ulitsky and Bartel, 2013). Nevertheless a set of functional LncRNAs with diverse functions and even more diverse ways of functioning have been identified. Recent studies (Ponting et al., 2009; Ulitsky and Bartel, 2013) have highlighted the diverged repertoire of mechanisms through which LncRNAs can regulate gene expression directly or indirectly, in cis or in in-trans. Most of the functional LncRNAs are multi-exonic but two of the most well known ones are long mono-exonic transcripts. Malat1 is involved in

the organization of nuclear speckle domains and Neat1 is essential for para-speckle formation, its expression is also induced in mouse brains after infection from Japanese encephalitis virus. Hotair is transcribed from the HOXC gene complex but regulates and silences the expression of the HOXD genes which lie on a different chromosome. All those LncRNAs are examples of functional in-trans transcriptional regulators.

Some very well known LncRNAs are found to regulate gene expression in-cis. Evf2 regulates Dlx5 and Dlx6 genes involved in neuronal differentiation. Xist is essential in the inactivation of X chromosome in eutherian mammals. It functions in a similar fashion to Kcnq1ot1, Air and Nespas which are necessary for the epigenetic gene silencing of imprinted genes in their clusters which have close proximity. They are all long transcripts with very little splicing and long exons. LINoCR activates the LYZ gene by its transcription. Its transcription *per se* activates an upstream enhancer by unbinding CTCF from a local insulator. PTENP1, which has derived from accumulated mutations after a duplication event of the PTEN known onco-gene, regulates the expression of PTEN and loss of function mutations are associated with cancer. Pbcas4, which is conserved in rodents, is predicted to act as a competing endogenous RNA where it antagonises with certain mRNAs for binding to specific miRNAs. SRG1 acts through transcriptional interference. Its transcription through the promoter of the SER3 gene suppresses the ability of the promoter to initiate transcription of the protein coding gene. Moreover a set of LncRNAs upstream of the fbp1 gene when transcribed change the chromatin structure and allow for the transcription of the fbp1.

In addition a set of protein binding LncRNAs bind to transcription factors, like LncRNAs upstream of the CCND1 gene and upstream of the DHFR gene, and suppress expression of protein coding genes.

A special class of LncRNAs is the antisense LncRNAs, which lie on the opposite strand of protein coding genes. Antisense LncRNAs can overlap some of the genomic regions of the gene model on the opposite

strand to various extents and they can regulate its transcription. Progression of transcription of a pair of protein coding gene and antisense LncRNA can be convergent or divergent. Furthermore LncRNAs can be regulated by a bidirectional promoter, i.e a common promoter with the sense protein coding gene in divergent transcription. Or they can be regulated by a polymerase II complex which transcribes the LncRNA convergently with the protein coding gene on the opposite strand (Albrecht and Ørom, 2016). In terms of regulatory characteristics, transcription of a LncRNA can activate the transcription of a protein coding gene on the opposite strand (Albrecht and Ørom, 2016) or it can silence by producing a ncRNA which acts as a competing endogenous RNA for the protein coding gene's mRNA (Han and Jan, 2013).

Known pain-related LncRNAs

Recently a number of studies have been published regarding the functional repertoires of LncRNAs in the context of pain. In 2013 Zhao et al (Zhao et al., 2013) identified a natural Kcna2 antisense LncRNA in DRG neurons. This natural antisense transcripts has been identified in rat, mouse, monkey and human and suppresses the expression of the Kcna2 gene, a voltage gated potassium channel. The antisense LncRNA is upregulated in peripheral nerve injury that causes neuropathic pain. On the other hand the protein coding Kcna2 is downregulated after nerve injury in mice. That distinct opposite expression pattern and the fact that most of the antisense Kcna2 overlaps most of the region of the protein coding gene suggest that the LncRNA regulates in cis the expression of Kcna2. This KCNA2 antisense LncRNA is the only known LncRNA that is proven to be functionally involved in neuropathic pain. It is expressed only in DRG tissue, in rat, mouse, monkey and human. It is also very lowly expressed, compared to Kcna2 mRNA. The main factor that induces the Kcna2 antisense LncRNA after nerve injury is the MZF1 transcription factor, belonging to the family of zinc finger proteins (Marie Lutz et al., 2014). This LncRNA can be an endogenous trigger in neuropathic pain as nerve injury induces its expression, which in turn reduces the expression of Kcna2

mRNA resulting in an increase of ectopic activity in large and medium DRG neurons (Marie Lutz et al., 2014).

In 2016 Wu et al (Wu et al., 2016), analysed the whole transcriptome of rat's DRG after the SNL pain model. The identified a huge amount of significantly DE genes (about 30%) and also they identified 944 non-coding RNA significantly DE. These ncRNAs are mainly intergenic RNAs and antisense RNAs.

In 2015 Jiang et al (Jiang et al., 2015) identified 511 differentially expressed LncRNAs in the spinal cord of mice after the spinal nerve ligation neuropathic pain model. 35 of those LncRNAs had neighbouring or overlapping significantly differentially expressed protein coding genes.

In 2015 Koenig et al (Koenig et al., 2015) identified a natural SCN9A antisense transcript in DRG conserved in human and mice. This antisense LncRNA is expressed in similar tissues like SCN9A, a gene coding for one part of the Nav1.7 sodium channel. The antisense LncRNA has long overlapping regions with the protein coding gene and down-regulates the gene on the opposite strand. However the authors report that there is no significant DE of the gene nor the antisense LncRNA in animal models of pain.

Overview of computational pipelines for identifying LncRNAs

As more and more studies for LncRNAs are getting published and more LncRNAs are being identified there is an emerging need for data repositories and standardized computational pipelines regarding LncRNAs. Currently there are 24 distinct, published annotation pipelines for the identification of LncRNAs (Xu et al., 2016). The majority of them are independent studies, focused on a specific organism or a specific biological process and dealing only with one kind of high throughput molecular biology data. There are also pipelines developed as part of more complex gene annotation consortia like NONCODE (Xie et al., 2014), LNCipedia, GENCODE (Harrow et al., 2012). 15 out of those 24 annotation resources use RNA-seq as the main data source. More specifically the majority of

them use *ab initio* assembly, where reads are first mapped to a reference genome, paired end-reads, a minimum length threshold of 200bp, a filter which filters out mono-exonic transcripts, a coverage threshold and a method to assess coding potential. Some studies do incorporate other sources of data like ChIP-sec and/or epigenetic signals. In terms of databases there are 15 distinct publicly available databases of LncRNAs which are populated by data annotated mainly for the computational pipeline briefly presented above. Xu (Xu et al., 2016) and Illot and Ponting (Ilott and Ponting, 2013) have published very comprehensive reviews regarding those computational resources. In this study, in chapter Methods for identifying LncRNAs and analyse RNA-sequencing data we present a computational pipeline that identifies novel LncRNAs using RNA-sequencing data. The main advantage of the proposed pipeline is that it has been specifically developed for the identification of novel LncRNAs and not all novel isoforms. Thus its is efficient and not computationally expensive and more importantly it does not rely on the FPKM metric to assess transcription strength, but it rather produces predicted gene models of LncRNAs suitable for downstream analysis with the proper framework regarding calculation of differential expression.

RNA-Sequencing

RNA-seq is a next-generation sequencing, or sequencing by synthesis method. It is used both for Differential Expression (DE) analysis and to identify unknown genes or non-coding genomic regions i.e. transcribed loci that do non encode for proteins (Trapnell et al., 2010; Weikard et al., 2013). Due to recent developments in next generation sequencing and particularly in RNA sequencing, we are now able to identify genes and previously un-annotated transcripts, such as LncRNAs, and to estimate their expression levels at the same time (Cabili et al., 2011; Young et al., 2012). These Long non-coding transcripts, which might have a regulatory role in gene expression, are the focus of much recent research

which aims to identify them in different species and infer their functional role.

Briefly RNA-seq involves the conversion of RNA, in whole or an enriched subset of RNA, to a library of complementary DNA (cDNA) fragments with specific strings of nucleotides, i.e. adaptors, attached at both ends. Then these fragments are amplified and each molecule is sequenced by synthesizing the complementary sequence from one end in single-end protocol or from both ends in paired-end protocol. This is a high-throughput process which generates numerous short reads with lengths varying from 50 to 400 bp (Wang et al., 2009).

Following the amplification and sequencing steps RNA-seq involves the *de novo* or *ab initio* assembly of multiple short reads of cDNA. The transcriptome can be assembled and analysed by sequencing those double stranded cDNA fragments reversely transcribed from RNA in the case of the *de novo* assembly, or reads could be mapped onto known genome assemblies in the case of the *ab initio* assembly also known as a *map-first strategy*. *Ab initio* assembly can reduce the computational burden of transcriptome reconstruction and it can produce more accurate results.

Thus, one of the obvious advantages of this high throughput sequencing technology, by contrast to micro-array probes, is that it does not depend on an *a priori* knowledge about the transcriptome and thus makes possible both the identification of yet unknown gene models or the *de novo* genome assembly for organisms with yet unknown genomes (R. Li et al., 2009; Mak, 2011). Most RNA-seq studies rely on *ab initio* analysis methods, where reads are first mapped on a known genome assembly.

RNA-seq analysis usually involves counting and mapping of the numerous short reads. Hence it is considered a count-based method. When used for detecting fold changes in expression of RNA between different conditions RNA-seq can quantify changes over 8000 fold with low background noise. It can detect different isoforms, splicing events and lowly

expressed transcripts (Wang et al., 2009). An overview of the RNA-sequencing procedure is provided in Figure 6.

RNA isolation and library construction

In order to carry out RNA sequencing we need to extract RNA from tissues of interest and produce a sequencing library. Library construction involves several steps and manipulations and is generally considered to be the step in RNA-seq which introduces most of the undesired variance (Seqc/Maqc-Iii Consortium, 2014).

The two most common methods are phenol extraction and column purification, but a lot of labs have developed hybrid protocols combining these two methods. A hybrid method of combined phenol extraction and column purification can be used with excellent results (Bartus et al., 2016). According to this method two steps of RNA isolation and extraction are carried out, resulting in sufficient yield of high quality pure RNA. Tissue is first homogenized and then mixed with chloroform following the phenol extraction method. After centrifuging, the rest of the aqueous liquid containing the nucleic acids, which stays on the top of the tube, should be removed. This solution is then subjected to the column purification method. RNA is then extracted and all samples are subjected to on-column digestion in order to prevent genomic contamination i.e. presence of DNA in the RNA samples.

The concentration of RNA in the samples is first measured using a nanodrop and if quantity and quality of samples is found to be good the sample is sequenced. Then a sequencing library of complementary DNA (cDNA) is constructed using a stranded or un-stranded protocol. The strand-specific (deoxy-UTP strand-marking protocol) dUTP protocol (figure 7) is the leading protocol for strand-specific synthesis of cDNA. The main features of the dUTP strand specific protocol are the incorporation of deoxy - UTP when synthesizing the second strand and the subsequent destruction of the remaining uridine - containing (dNTP) strand in the library. Thus after synthesizing the first strand of cDNA, dNTP is removed and the second

strand cDNA is then synthesized using dUTP. In this way the polarity of the transcripts is revealed (Parkhomchuk et al., 2009). In dUTP protocol the reads produced have reversed strandedness/direction, thus this should be taken into account for downstream analysis. Namely strandedness of reads should be reversed before counting.

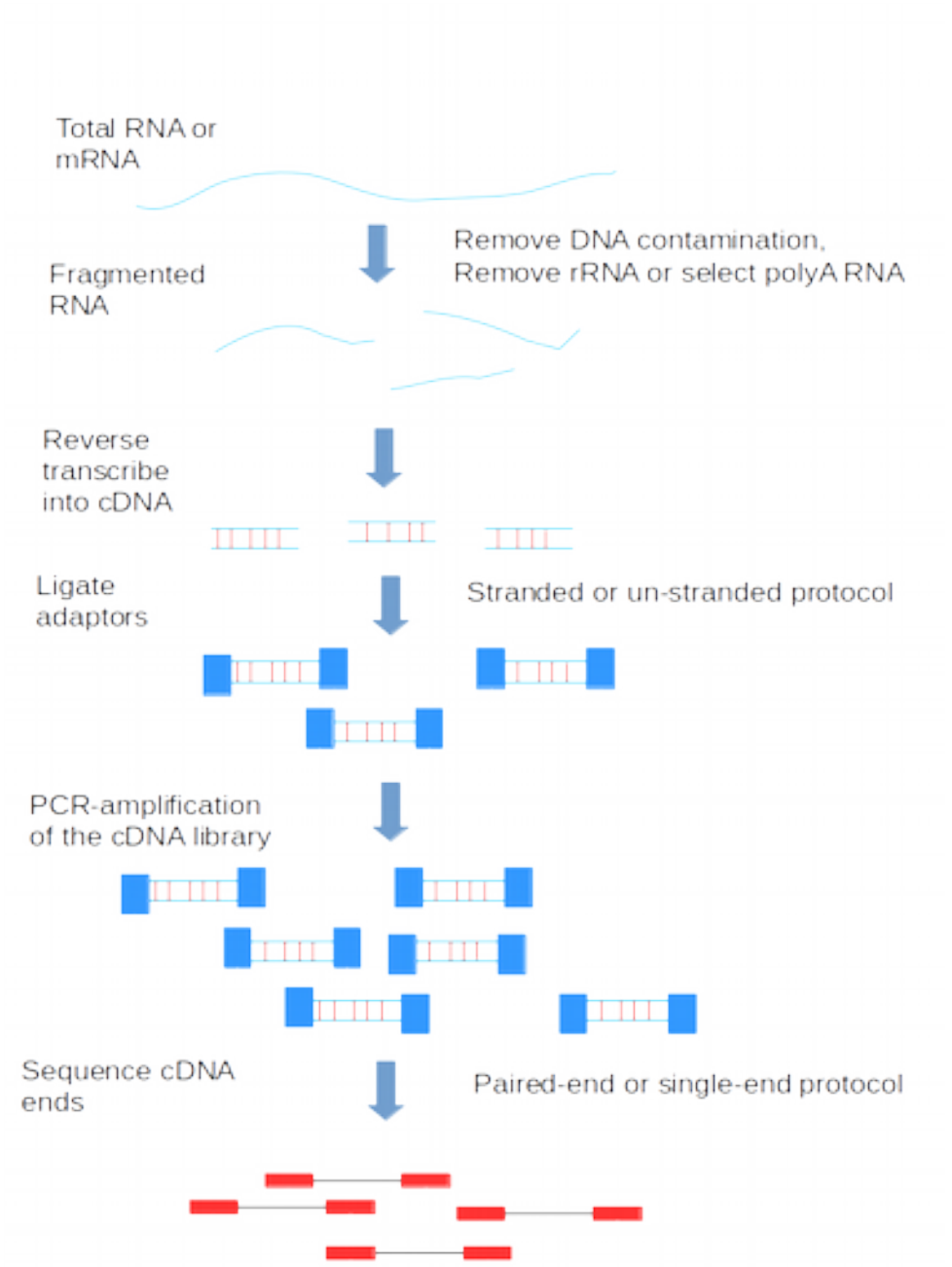


Figure 6: Overview of RNA-sequencing

Total RNA or mRNA (blue line) is purified and ribosomal RNA is depleted or poly-adenylated RNA is selected. The selected fraction of RNA is fragmented and then reverse transcribed into cDNA and amplified. Afterwards, sequence adaptors (blue rectangulars) are attached to fragments, the protocol could be stranded or un-stranded. The library of cDNA is PCR-amplified and then sequencing starts from one end or both ends.

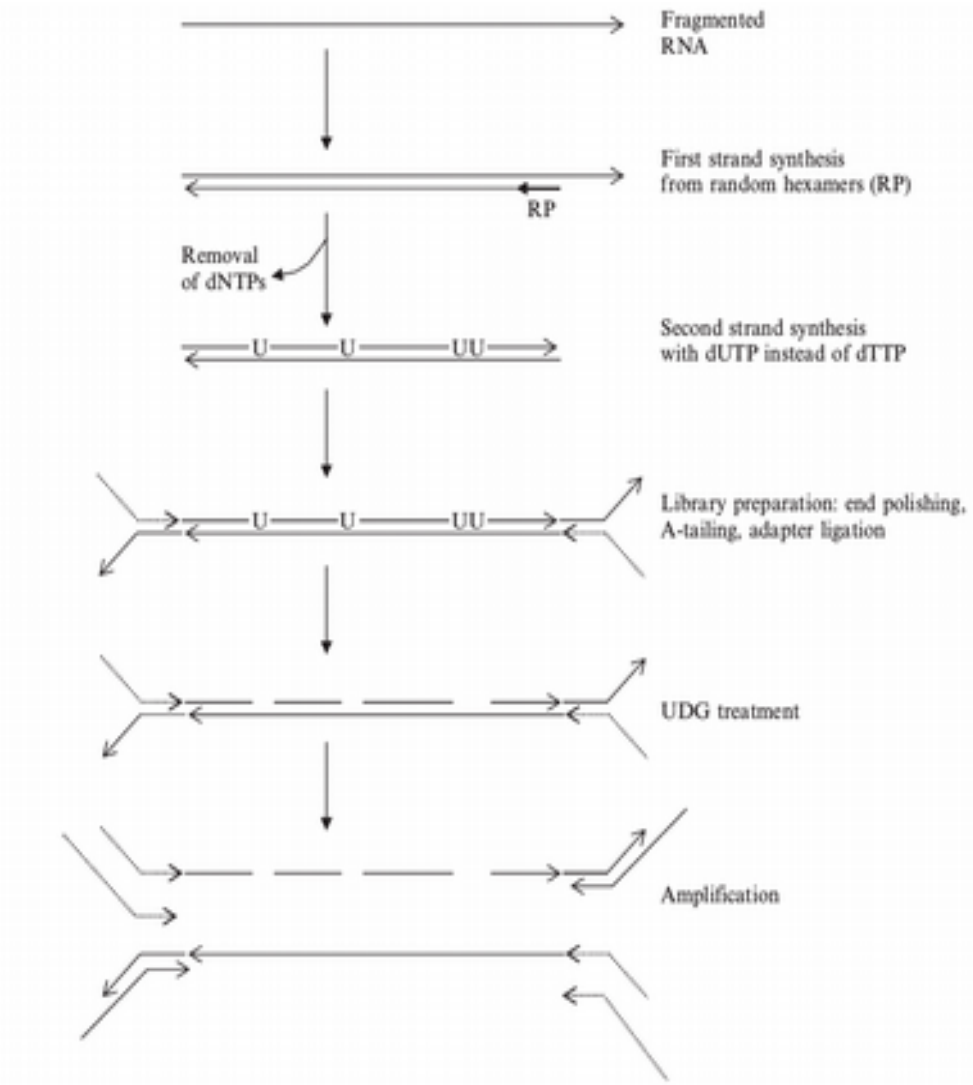


Figure 7: Overview of the dUTP strand specific protocol (Parkhomchuk et al., 2009)

Potentials and drawbacks

Despite its obvious benefits, RNA-seq has some drawbacks which can create noisy results. Most of the noise is usually introduced in the library preparation step but there are also significant bioinformatic challenges. First, the reads cannot always be unambiguously mapped to a certain genomic region due to repetitive sequences or erroneous base calling. On the other hand, some reads cannot be mapped at all because of certain differences in the sequence between the sample and the reference genome. This problem can become more severe in less well-annotated

genomes, like the rat genome, with a lot of unknown SNPs and with a poor estimation of the genomic variation in general. Gapped aligners like Top-Hat 2 and STAR (Dobin et al., 2013; Trapnell et al., 2012) can overcome this particular problem by allowing sub-strings of the read to be optimally mapped on the genome. This approach can facilitate either reads falling into splicing junctions or reads having differences in certain bases compared to the reference genome assembly.

Another problem, which could lead to overestimation of the expression of genes is that if a few genes in the sample are too abundant then they could use all the available reads of the given read depth. In addition small variances in the counts of lowly expressed genes, i.e. genes with low counts can produce very high log fold changes and gene counts cannot be directly comparable if they are the product of unequal sequencing libraries. These problems have led to the development of several normalisation methods that usually shrink the fold changes in some way. These methods usually scale the counts of each gene or genomic region by an estimation of the specific library size and estimate log fold changes by moderating them according to the general trend of changes observed in genes of similar expression strength.

We should note here that there are various methods for the analysis of RNA-sequencing data. One is the normalised count based approach (Anders et al., 2015; Love et al., 2014) where read counts represent relative transcript abundance for the scope of downstream differential expression analysis and another is an approach where read counts are normalised by the transcript length, the read length and the library size, i.e. total number of mapped reads, and transformed into a score named Fragments Per Kilobase of exon per Million mapped reads (FPKM) or Reads Per Kilobase of exon per Million mapped reads (RPKM), representing expression strength (Trapnell et al., 2012).

Additionally, some early studies showed that for RNA-seq read depth is much more important than biological or technical replicates

(Seqc/Maqc-Iii Consortium, 2014). Recently, more advanced data analysis approaches show that biological replication gives us significantly more experimental power than increasing the read depth above a threshold (Liu et al., 2014). Additionally biological replicates drastically reduce intra-sample variation and marginalise systematic biases (Seqc/Maqc-Iii Consortium, 2014). A comprehensive assessment of RNA-seq also reveals that most of the biases are not due to specific sequence characteristics, like the transcript length or the GC content of a particular genomic region. Rather, most of the data variability is attributable to library preparation (Seqc/Maqc-Iii Consortium, 2014; Xu et al., 2016).

Combining this with the sources of technical variation presented above, we can conclude that RNA-seq cannot be considered a robust method for estimating absolute transcript abundances, but rather a method for quantifying relative abundance differences, namely Differential Expression (Łabaj and Kreil, 2016; Rapaport et al., 2013). Moreover RNA-seq is considerably better than microarrays when it is carried out with a sequence depth of 50 million reads or more. A recent review (Perkins et al., 2014) comparing RNA-sequencing to exon arrays, when assessing transcriptional changes in the context of the SNT pain model in rat, showed that, although both methods agree in the genes they found to be significantly differentially expressed, RNA-sequencing clearly outperforms exon arrays. It has much better sensitivity in detecting lowly expressed genes, an attribute which is increased by increasing the read depth; a significantly higher dynamic range of fold changes and the ability to interrogate a much larger proportion of the genome and also to detect novel genes. All these desirable features can be extended by increasing the read depth and adding biological and technical replicates. Although, in terms of consistency between experiments carried out in different labs, RNA-sequencing is inferior to microarrays if we consider only genes with higher log fold changes (Seqc/Maqc-Iii Consortium, 2014). In addition, interestingly one of the most widely used RNA-seq data analysis pipelines, mostly due to its simplicity and ease of use, the “tophat 2 + cufflinks” pipeline is found to

show the worst consistency amongst all other RNA- seq pipelines and microarrays. This pipeline is heavily relying on the transcription strength FPKM / RPKM score paradigm. In general the count based approach is better suited for differential expression analysis while the FPKM/RPKM approach can be used to estimate absolute transcript abundance in a tissue or cell line (Anders et al., 2015; Love et al., 2014; Seqc/Maqc-Iii Consortium, 2014).

Statistical analysis of RNA-seq data is not trivial and requires biological replicates, normalization and gene filtering. These findings also suggest that DE results should be prioritized not solely by p.value but also by the size of the effect, namely by fold changes between samples. Other bioinformatic challenges arise from the sheer amount of data RNA-seq can produce and also due to the increasing complexity of the gene models to which reads should be assigned.

Analysing RNA-sequencing data

This study involves analysis of RNA-seq datasets associated with pain and particularly neuropathic pain. In this section we will briefly describe the main steps for analysing RNA-seq count data. This paragraph will give an overview, of steps which will be covered in more detail in the following chapters.

After mapping raw reads to a reference genome using an aligner we can acquire counts for specific features. A feature is represented by a genomic interval, a chromosome, a start and end position and the strand. Usually features are protein coding gene models and non coding transcripts like gene models of novel LncRNAs. Genomic features, like genes, are comprised of exons, introns and untranslated regions (UTRs). Different combinations of these, produce numerous transcripts arising from the same gene. As exons are the features of a gene model that are part of the mature RNA, after the introns have been removed during splicing, we calculate relative gene abundance by calculating the number of RNA-seq reads (or

pairs of reads when dealing with paired-end sequencing) overlapping its respective exons.

The aim of the study involving RNA-sequencing data analysis also dictates what might be the best method for assigning a distinct numerical score to those gene models in order to quantify transcript abundance. As we briefly discussed above, there are two broad families of methods. One is more appropriate for identifying absolute expression strength and the other is more appropriate as a first step in an analysis aiming to identify significant changes in relative transcript abundance, i.e. differential expression.

The first method calculates a score, usually normalised reads or ffragments per kilobase of transcript per million mapped reads using all available reads, which ideally should represent the number of cDNA fragments found in a sample, while the second method retains only reads which carry information in order to accurately assess statistical significance and log fold change between different conditions using a table of counts. We used the latter read-filtering and count-based approach as we were interested in identifying genes significantly differentially expressed between conditions of interest and not in performing an absolute quantification of expression strength, an elusive task as we discussed above.

Counting RNA-seq reads for DE analysis

A count-based RNA-seq analysis involves two steps. First, in the counting step a table of counts is calculated, which stores the raw number of sequencing fragments (single reads or pairs of reads) that overlap each feature of an annotation for each sample. Then, reads are normalised and statistical analysis quantifies the expression differences between conditions of interest and assesses whether differences in the amount of reads between different conditions reach statistical significance.

Counting reads for DE analysis needs specific approaches as repetitive sequences, wrong base calling, overlapping and nested gene annotations are making the assignment of RNA-seq reads on features an

inherently non-trivial task (Anders et al., 2015). For example if a set of reads can be assigned on the same time in two overlapping or adjacent loci we cannot efficiently compute differential expression analysis as we are counting the same signal twice and in the case of relatively short transcripts, like most LncRNAs, a very small number of reads could give an artificially high FPKM score. Moreover, if we have two or more loci with identical sub-strings of sequences, like in the case of paralogous genes, then we need to identify which one is actually differentially expressed. Additionally, the varying quality of the reference genome and the varying complexity and quality of the annotation set can lead in significant biases if we allow ambiguously mapped reads to contribute to the estimation of the relative expression abundance. In the case of computationally predicted annotations, like sets of novel LncRNAs, it is preferred to use a more conservative approach in counting features to overcome biases arising from the putative inferior quality of a computationally predicted annotation. That is why we used a conservative count-based approach, where reads assigned to a feature are being counted only once, but in combination with a strategy to retain some ambiguously mapped reads.

There are three approaches for counting features as the first step for a differential expression analysis (figure 8). The first method, which is the most conservative one, considers only the union of the un-ambiguously / uniquely mapped reads assigned on the group of exons of an annotated feature. All reads mapping to more than one feature are discarded. The second, “intersection strict” approach only considers reads fully contained in an exon, but in the case of reads mapping to more than one feature it assigns the read only to the feature which fully contains it. The third, “intersection not empty” method first calculates a disjoint version of the annotation consisting of only non-overlapping features while keeping track of all features originally found in the annotation. Then, when a read overlaps one of those non-overlapping features, which represent an original annotation feature, is being counted. Only reads that are fully contained in two originally overlapping features are discarded as ambiguous. Thus reads

that could be, even partly, assigned on a diverged part of two features with otherwise similar sequences they can be uniquely counted. This means that even if a fraction of a read overlaps two features, if there is another fraction of the read that uniquely overlaps a fraction of one of the two features then it would be assigned to the feature it uniquely overlaps partially. Because the third strategy allows for some ambiguously mapped reads to being counted (only once) to the feature they are the most likely to have derived from, it is better suited to the differential expression analysis of computationally predicted features.

Subsequently an RNA-sequencing analysis will typically look for significant differences in the expression levels of genes (features) between conditions of interest. The attributes of count data including its discrete nature, non-normality and dependence between the variance and the mean, must be taken into account. Natural variance within groups or conditions, technical biases, very lowly or very highly expressed genes and differences in library sizes can significantly affect the relative expression of a feature across conditions. Moreover high dynamic range and outliers, given a limited number of replicates, produce additional challenges.

Most of the methods for analysing RNA-sequencing data address these problems by not treating each gene separately but instead by borrowing information and constructing assumptions for gene expression using different genes in the same experiment (Anders and Huber, 2010).

Identification of significantly DE genes involves normalization to the library size, data transformations, an estimate of dispersion of genes between and within conditions and performing statistical inference using hypothesis testing.

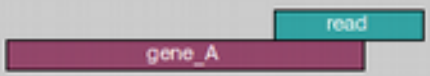
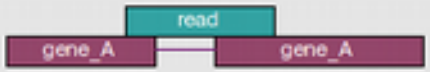
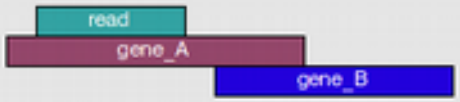
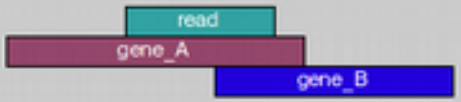
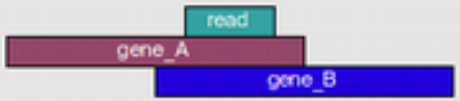
	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Figure 8: Count modes of RNA-seq reads, different strategies assign different reads to gene A and gene B. Image courtesy of (S. Anders et al., 2015)

Differential Expression analysis for count-based RNA-seq data

Typically, the Table Of Counts (TOC) produced in the counting step is supplied into an algorithm which then estimates fold changes between conditions, filters out outlying features with artificially low or high counts or low or high dispersion, and assesses statistical significance. In this study we used DESeq2 (Love et al., 2014) the latest and most widely used R / Bioconductor (Gentleman et al., 2004) method for analysing RNA-seq count data. DESeq2 and the method we used for counting reads, which is HTSeq (Anders et al., 2015), are tightly coupled and highly compatible.

DESeq2 estimates log fold changes of genes and infers statistical significance of the observed change. To accurately estimate log fold changes (lfc) and to overcome biases introduced due to the diverge average expression and the dependence of variance to the mean, DESeq2 borrows information across all genes in an experiment. Then moderates the dispersion of genes through an empirical Bayes shrinkage method based on the assumption that genes with similar average expression strength have similar dispersion. Additionally, it shrinks log fold changes of genes towards zero according to how much information (counts, dispersion, degrees of freedom) is available. Furthermore, it filters out outlying genes which show artificially high or low dispersion or counts. The table of content entries that report the number of RNA-seq fragments assigned to each gene for each sample are modelled using the negative binomial distribution. The mean of the distribution is then scaled according to a normalisation factor proportional to each sample's library size. Subsequently, the within conditions variability or dispersion is modelled by a dispersion parameter. Genes of similar expression levels, i.e. similar mean of the negative binomial distributions are assumed to show similar dispersion. First each gene is treated independently, and dispersion is estimated using the maximum likelihood criterion. Then this estimation is normalised to accommodate for dependence on the average gene expression strength by fitting a smoothed curve. Finally the dispersion estimates for each gene are shrunk towards the predicted smoothed dispersion curve. The strength of

normalisation / shrinkage depends on the degrees of freedom, namely the sample size and the number of replicates and on an estimation of how far actual dispersions are from the fit.

As genes with low counts are more sensitive to small variations and thus inherently noisier than genes with relatively hight counts, DESeq2 shrinks log fold changes for those genes according to the amount of information available using an empirical Bayesian approach. Genes with low counts or high dispersion in experiments with few degrees of freedom (small sample size, few replicates) undergo stronger shrinkage. These normalised and transformed log fold changes are then used for hypothesis testing.

Hypothesis testing

To estimate differential expression between conditions we use a Generalized Linear Model (GLM) formalism. For every transcript considered, raw counts are modelled as following a negative binomial distribution, as is standard practise for RNAseq counts. The mean of that distribution is linked to a linear predictor via a link function, in this case of logarithmic form. In the simplest form, the GLM contains only one term reflecting the status of a sample (condition versus control) and the GLM fit returns a coefficient corresponding to the log2 fold change in expression between the two types of samples. The use of GLM models also allows for more complex designs, where a number of independent variables can be included in the linear predictor to reflect additional covariates affecting gene expression.

Then typically we test if the estimated coefficient calculated using the correct contrast is equal to zero, i.e. the null hypothesis. Usually this involves classical hypothesis testing. There are a number of tests for classical hypothesis testing like the Student t-test, Fisher exact test or the Wald test. Typically these tests compare two distributions and answer the question whether the values observed are too extreme to belong within a distribution of reference. The process is as follows: the coefficient estimate,

the normalised fold change between conditions is divided by its variance producing z-scores which are then compared to a normal distribution.

First we should precisely specify a null hypothesis (usually no change-coefficient is zero) and an alternative (usually coefficient is not zero). We then use the sample data assuming that the null hypothesis is True and because of that we can calculate the test statistic using the mean and the standard variance of the observed data. Consequently, we can calculate the p.value for a specific test by using the known distribution of the test statistic.

The p.value is a statistic that describes: "*If the null hypothesis is true what is the probability that we would observe such extreme measurements*". Then, typically, we set an arbitrary cut-off threshold which represents the probability that we would reject the null hypothesis when it is actually true. This is called type 1 error or false positive. Similarly, type 2 error is when we do not reject the null hypothesis when it is not true. These errors are meaningful in the context of specificity for type 1 errors and sensitivity for type 2 errors. Ideally we would like an optimal trade-off between sensitivity and specificity. We should stress attention again to the fact that p.values represent how likely it is to observe our data when we assume that the null hypothesis is true. Thus the actual probability of a test to be erroneous cannot be directly inferred by the p.value, but we rather need to take into account the actual chance of a real effect. In our case the real effect of a gene to be significantly DE between certain conditions (Nuzzo, 2014).

As we are dealing with thousands of genes it is reasonable that errors will be aggregated by the number of tests we perform. For example if we allow for a probability to reject the null hypothesis when it is actually true to 5%, or we seek for a p.value < 0.05 in order to reject our null hypothesis in one test, then we can have our null hypothesis being rejected 5 times if we perform 100 tests. Thus we use only adjusted p.values for multiple testing, where the simplest way to adjust is to divide the cut-off threshold by the number of tests. Although this produces a consistent threshold it

dramatically affects sensitivity, thus more usually we use more sophisticated methods like the False Discovery Rate (FDR), where we control the total number of false positives over the whole set of comparisons to be below a certain threshold. In the current study, in most cases, we rejected the null hypothesis only when we observed adjusted p.values < 0.05.

Explain complex interactions of many variables

In most of the studies presented in this thesis we are dealing with large datasets having too many variables and a relatively small number of samples. In general we are dealing with problems that cannot be analytically solved with certain systems of equations. Moreover we usually have highly heterogenous data, where different variables have very different scales and variances and different types of variables (numeric, categorical, ordinal) co -exist. Typically we would like to infer, or perform a reductionist approach, on these datasets in order to either identify the most important attributes that characterize certain conditions, or to identify coherent sub-populations with specific properties in our samples. These problems could be formulated into problems of identifying the most contributing set of coefficients in a linear model, i.e. finding the most powerful predictors, or finding the combination of predictors that could effectively explain most of the data’s variance or optimally classify our samples in a biologically meaningful way.

Principal Components Analysis and varimax rotation

After normalization a dataset is ready to undergo some form of reduction in dimensionality. This is exactly the process that reducts the variables or combinations of variables with the largest explanatory power. A standard well known method is the principal components analysis (PCA). Principal components (PC), i.e. linear combinations of the original variables, which are orthogonal with each other and explain the maximum possible variance of the original data, are calculated and those that cumulatively explain more than a certain fraction of the original data's variance are retained. Principal components can be less or at most equal to the number of original variables. There are several methods to decide the

optimal number of principal components. One method uses scree plots that show the number of PCs against the eigenvalues and one can select the number of components that corresponds to an elbow in the graph, i.e. after that point the rate of the reduction in the eigenvalue's values is getting significantly lower. Or one can select the PCs that explain more than a certain fraction of the original data's variance, typically 80% of the variance, or just select only principal components with eigenvalues of 1 or more. Plots of the two first principal components are used in order to examine how those PCs can effectively partition data and when the only purpose is this kind of visualisation only two components are used. We used the latter technique for visualising sample separation based on the two first principal components of a gene expression matrix. We should note that PCA is very sensitive to the original distribution of the data's variance, thus it requires careful normalisation (scaling) of the data.

PCA is essentially a singular value decomposition of a data matrix (*Principal Component Analysis*, 2002). In PCA, a covariance matrix is split into a scalar part (eigenvalues) and a direction part (eigenvectors). Loadings are the eigenvectors, the coordinates of the variables for each principal component (dimension), divided (scaled) by the square root of their respective eigenvalues. Thus loadings are the most efficient way to observe the normalized contribution of the original factors and variables into the distinct principal components ("FactoMineR," n.d.).

In PCA, principal components, which are essentially projections of the data on the directions defined by the eigenvectors, are uncorrelated. In other words the space of the principal components is an uncorrelated orthogonal basis set of the original data. The eigenvectors are in turn orthogonal directions. As loadings are the eigenvectors, if we select a certain number of principal components we can observe how different factors contribute on their respective loadings. Thus it will be easier if we could have as sparse components as possible, with some few factors with relative high values and with a lot of factors with zero or near zero values. In this study we used the varimax (Kaiser, 1958) rotation in order to rotate the sub-

space of the orthogonal basis of the loadings and increase their ability to optimally separate data. Varimax rotation is taking place on the latent space, the space of loadings and not the space of the original principal directions. Namely, it rotates the sub-space of the orthogonal basis of the loadings by multiplying them with a $\boldsymbol{K \times K}$ square orthogonal matrix $\boldsymbol{T}$, which is calculated as the orthogonal square matrix which can carry out the orthogonal decomposition of the matrix which holds the first $\boldsymbol{K}$ ($\boldsymbol{K} < nr$ of original loadings) loadings. The aim is to make the matrix of the rotated components as sparse as possible.

Overview of thesis chapters

The next chapter, **chapter 2: Methods for identifying LncRNAs and analyse RNA-sequencing data,** of this thesis presents an automated bioinformatics pipeline, aiming to identify novel LncRNAs using RNA-seq data, to calculate differential expression, to annotate these LncRNAs according to their genomic context and to perform functional enrichment of the known DE genes. This customised pipeline integrates a strategy for the identification of LncRNAs which exploits the previously un-annotated transcribed areas found applying a coverage threshold to the RNA-seq data, alongside with un-annotated splicing junctions *de novo* identified by a gapped aligner. This pipeline has the advantage of producing gene level annotations of LncRNAs suited for downstream counting and analysing for differential expression. The next two chapters present results from using this pipeline

Chapter 3: Transcriptional changes of protein coding genes and novel LncRNAs in rat's DRG, presents a rubric for analysing RNA-seq data using the customised pipeline presented in **chapter 2** and results from rat dorsal root ganglion (DRG) under the spinal nerve transection (SNT) pain model. This study investigates the transcriptional changes of both known genes and novel LncRNAs.

Chapter 4: Transcriptional changes in DRG of two mouse strains experiencing high and low induced hypersensitivity, further investigates

transcriptional changes under animal models of peripheral neuropathy. In this chapter we present results from analysing RNA-seq data from two mouse strains DRGs under the spared nerve injury (SNI) pain model. Those two strains had significantly different induced hypersensitivity after the pain surgery and we carried transcriptional profiling investigating both novel predicted LncRNAs and known genes.

Chapter 5: Clustering of patients with diabetic neuropathy reveals distinct neuropathic pain dimensions, of this thesis presents another bioinformatic approach for the further understanding of painful neuropathy. This time we present a rubric and results from an exploratory data analysis of clinical data and data obtained from quality of life – pain questionnaires from patients suffering from diabetic neuropathy. Analysis of this data gave us useful insights regarding the correlation of various questionnaires with neuropathic pain intensity, the correlation of neuropathic pain intensity with clinical factors and the diverge somatosensory clusters of diabetic neuropathy that can be related to the reported intensity of neuropathic pain.

Finally, in **chapter 6** we provide a brief summary of the conclusions drawn from the thesis chapters and discuss immediate future work.

Methods for identifying LncRNAs and analyse RNA-sequencing data

Overview of computational identification and DE of LncRNAs

In this study we present methods and results of a bioinformatics pipeline developed to analyse RNA sequencing data, in order to computationally identify novel LncRNAs, which may be functionally important as they are differentially expressed (DE) between conditions of interest. As discussed in Introduction, numerous studies have identified LncRNAs using RNA-sequencing data. Almost all of them have used computational pipelines made for identifying novel isoforms of known genes (i.e. pipelines using Cufflinks (Trapnell et al., 2012)) and others have used computational methods for a complete *de novo* or *ab initio* identification of the transcriptome (i.e. methods using scripture (Guttman et al., 2010) and stringtie (Pertea et al., 2015)). All these methods can identify novel LncRNAs but they are either computational expensive and not very sensitive as they try to reconstruct the complete transcriptome, or they do not produce annotations compatible with count-based methods for DE analysis, as they just identify sets of transcripts and isoforms using the FPKM transcription strength estimation. In this chapter we present a method tailor-made for the identification of LncRNAs. Our method predicts complete models of sufficiently expressed novel LncRNAs, using a mapping-first strategy, which are in suitable form for downstream DE analysis. Thus we can identify novel LncRNAs and infer their function by analysing their expression profile under certain biological conditions of interest.

In the next chapter we present results from applying this pipeline to analyse RNA-seq data of rats under the Spinal Nerve Transection (SNT) pain model and in the following chapter we present results from RNA-seq data of mice which underwent the Spared Nerve Injury (SNI) pain model.

The aim is to identify putative LncRNAs which may have a certain functional role in neuropathic pain. This is inferred mainly by their expression pattern in pain relevant tissues under different biological conditions of interest, their genomic context and the abundance and consistency of their expression. We computationally predicted and annotated loci that express LncRNAs that were differentially expressed (DE) between conditions and strains related to neuropathic pain. We also assessed the general transcriptional changes under neuropathic pain models by detecting significantly DE known/annotated genes and by performing functional enrichment analysis on those gene sets.

The main advantage of this customised pipeline is its integration with the proper framework for analysing RNA-sequencing data in order to identify DE genes. Thus we propose an approach that identifies gene models of LncRNAs that are suitable for further downstream analysis, namely counting with a method that does not overestimate counts and deals with ambiguous reads, like HTSeq, and DE analysis with a count-based approach that does not overestimate log fold changes for lowly expressed transcripts, like DESeq2. That is not trivial, as LncRNAs are notoriously lowly expressed, which makes DE analysis difficult and error prone. Moreover summarisation at the gene level is very important, as the identification of novel transcribed loci / gene models of LncRNAs can give many slightly different transcripts, some of them overlapping, which can seriously affect downstream DE analysis. Thus our approach is significantly different to the approaches described in Introduction in the way that it does not identify a set of transcripts, using metrics suited to the transcript expression strength and not to DE analysis, but instead we identify predicted full gene models suited for further DE analysis.

More specifically this computational pipeline was designed in collaboration with experimental biologists and it can use data from a biologically relevant RNA-sequencing experiment in order to predict putative LncRNAs, calculate differential expression and perform functional enrichment. The aim is not to perform a complete and complicated

reconstruction of the whole transcriptome, but rather to develop a fast, simple and intuitive pipeline that could identify loci, i.e. genomic regions at the gene level, of putative LncRNAs which are the most probable to be functionally important and are expressed at levels that allow them to be validated in the wet lab in the context of a very specific biological experiment. Thus the pipeline gives us the ability to have a more complete image of transcriptional changes, both in annotated genes and un-annotated LncRNAs, under specific conditions.

Analysing several RNA-Seq datasets using this pipeline allowed us to computationally predict putative gene models of LncRNAs using RNA-seq expression data and *de novo* annotated splicing junctions. We have also studied the expression of these LncRNAs in the context of the expression profile of their antisense protein coding genes and genes functionally validated in animal models of pain found in “The Pain Genes Database” (Lacroix-Fralish et al., 2007).

Methods

The main task of this computational pipeline is to identify novel LncRNAs which could be functionally important and thus differentially expressed in a biologically relevant experiment. For this purpose we have used the statistical programming environment R (R Core Team, 2015), several Bioconductor packages (Gentleman et al., 2004) and customised unix scripts. A brief overview presenting all the main steps of the pipeline is in figure 1.

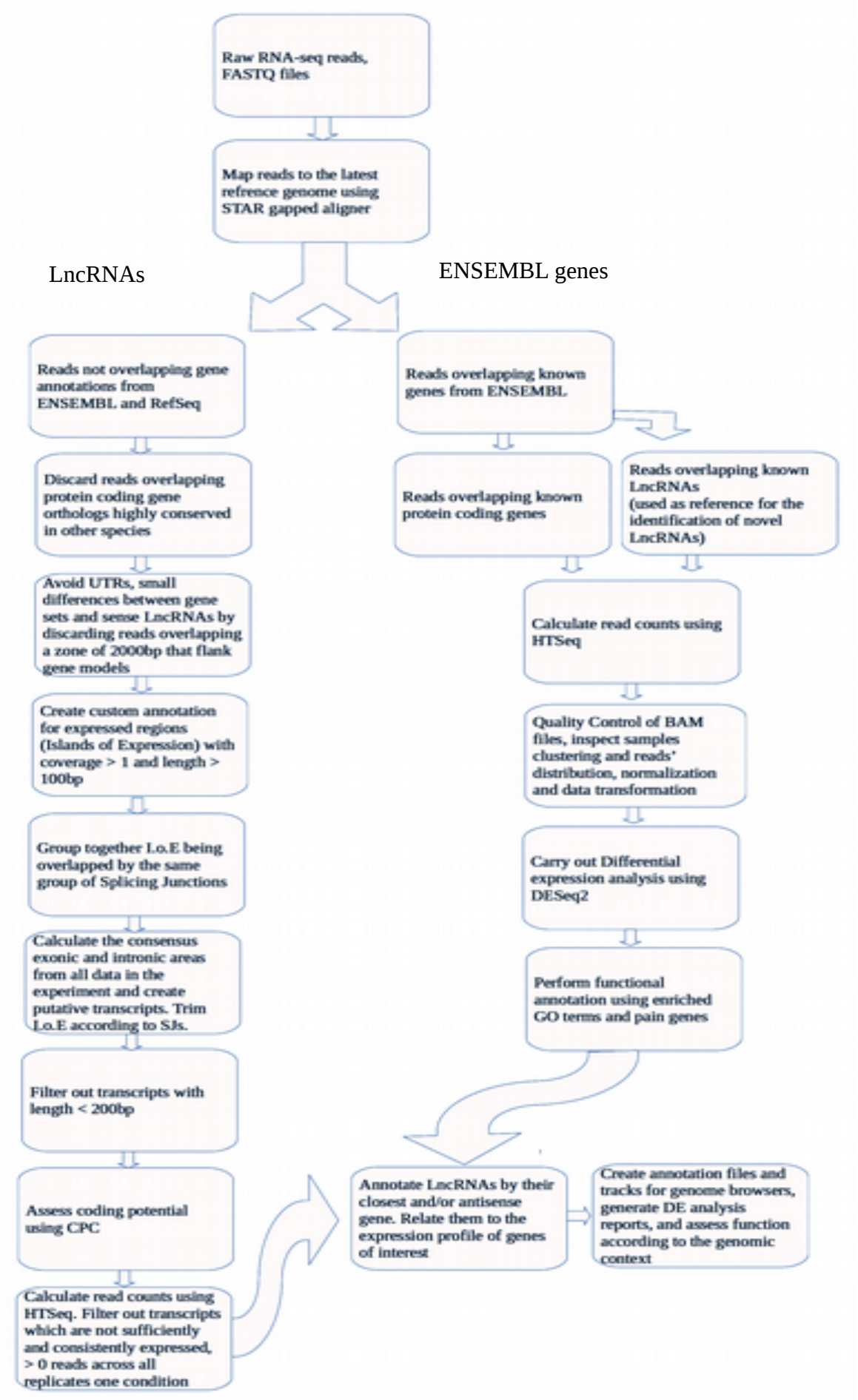


Figure 1: Flowchart of computational pipeline

RNA-Seq and library preparation

The main input of the pipeline is RNA-seq data from a specific biological experiment. Although all RNA-seq datasets could be analysed, there are some limitations and certain sequencing specifications will definitely produce much more accurate results.

Firstly, data should be obtained for at least two conditions of interest and there should be enough technical and biological replicates. Recent reviews (Seqc/Maqc-Iii Consortium, 2014) have shown that RNA-seq results are highly dependent on effects arising in the library preparation step. Thus in order to obtain an accurate estimation of within samples variance for every feature of interest we need at least two biological replicates for each condition and some technical replicates that will control the variance introduced in the library preparation step. Moreover, in most cases, pooled samples, i.e. when RNA from a certain number of different samples is put together to create one data-point, should be used as an effective way to marginalise unwanted variance within conditions. When technical replicates are available (i.e. different libraries from the same biological material) those are mapped separately to the genome and then the output of the mapping algorithm, i.e BAM files for each replicate, are merged. When biological replicates (i.e. different samples from the same condition) are available they are not merged but used to estimate within samples variance.

Also, a sequencing depth of more than 50 million reads per sample is required for the identification of lowly expressed transcripts and paired-end reads would allow for a significant increase in the accuracy of the identification of expressed loci ,as fragments of cDNA are sequenced from both ends (Ilott and Ponting, 2013; Perkins et al., 2014; Seqc/Maqc-Iii Consortium, 2014). As we wish to identify expressed regions outside known gene models, to reconstruct those transcripts our results are highly dependent on the read depth, the read length and whether reads are paired-end or single-ended. Genomic sequences of low complexity, complex gene models and the way of mapping short reads to the genome which introduces some inherent uncertainty makes the mapping of shorter single-ended reads

a more difficult problem that the mapping of longer paired-end reads. As the pipeline uses the output of a gapped aligner, the ability of that aligner to accurately align reads to the genome is crucial. Moreover, as we need to use novel splicing junctions outside known gene models in order to reconstruct gene models of putative LncRNAs, our results depend on the number and accuracy of the splicing sites identified by the gapped aligner. Longer reads have much more probability to fall into splicing sites and a higher read depth gives much more accuracy in detecting splicing events and contiguously expressed regions according to the number of reads that overlap them (Perkins et al., 2014). For these reasons we used reads of 100bp and longer, paired-end with a read depth of more than 50 million reads per sample. These sequencing specifications are not extremely high but are known to produce accurate and reproducible results (Seqc/Maqc-Iii Consortium, 2014).

Since there is a finite RNA-seq read depth and since certain fractions of RNA are much more abundant, we used enriched RNA for certain fractions of the transcriptome. The main goal is to discard the ribosomal RNA which would otherwise take up the largest part of our read depth. One way is to select only poly-adenylated RNA in order to have samples enriched in mRNA. The other way is to specifically remove ribosomal RNA from the sample, this process is called ribodepletion. In order to maximize the ability to identify novel transcripts it is advisable to use the ribodepleted fraction of the total RNA for further sequencing. Ribodepletion excludes the ribosomal RNA but it does not filter all other types of RNA which might be of biological interest as we are mainly interested on the non-coding part of the transcriptome.

Finally, in order to be able to identify antisense LncRNAs we needed a library preparation protocol which retains strandedness. It is also important to know the direction of transcription in the downstream analysis in order to validate targets in the wet lab. We used the dUTP protocol (see chapter Introduction, section RNA-sequencing), which synthesizes both the forward and reverse strand in the amplification step of library preparation.

Aligning reads to the genome

As we used the *ab initio* / *map first* assembly strategy, we first mapped raw reads stored in FastQ files to the reference genome. For this purpose we used the STAR aligner (Dobin et al., 2013) in its latest version. As we were looking for un-annotated parts of the transcriptome, it is crucial that we used an aligner that could accurately align short to medium reads by taking into account mismatches, insertions and deletions, repetitive sequences and wrong base calling and at the same time an aligner that performs well in aligning into genomic regions that are not continuous, but they have been joined together by splicing.

STAR is an aligner that has been developed and designed with all these features in mind. It makes use of the Maximum Mappable Prefix (MMP) concept, where, instead of aligning the whole sequence of an RNA-seq read, takes the longest substring of the read's sequence that exactly matches the sequence of one or more substrings of the reference genome's sequence, figure 2. The genome's substring with the maximum length sets the Maximum Mappable Length (MML). Then the unmapped part of the read is subjected to the same process of substring mapping. This process is iterated until no more of the read can be mapped. In this way, if for example a read falls on a splicing junction, the first seed will map to the donor site and the second seed to the acceptor site. Finally all seeds that are aligned in a predefined window, which intuitively defines the maximum intron size, are clustered and stitched together. By recurring passes of the MMP step, reads with mismatches can be aligned. Moreover, paired-end reads are clustered and stitched together as a single sequence in order to preserve the natural attribute of these reads to represent both ends of the same sequence / cDNA fragment.

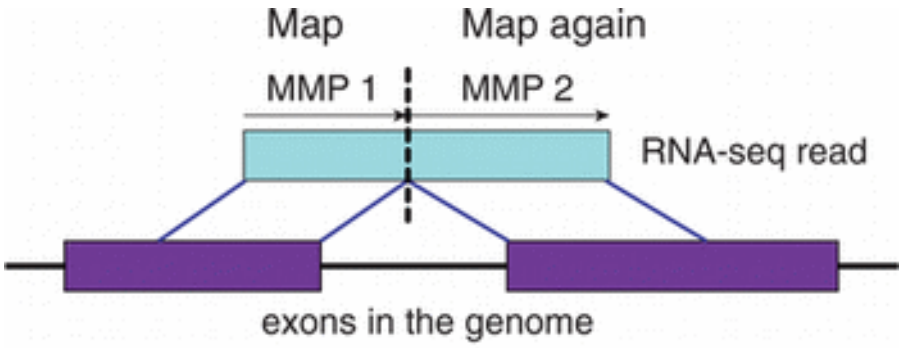


Figure 2: The first substring (MMP 1) that matches exactly a substring of the reference genome is first mapped to an exon, then a gap is introduced and the second substring is being matched on another exon. This read is assigned to the exons, i.e. produces a certain coverage, and at the same time identifies a splicing event. Image courtesy of (Dobin et al., 2013).

Additionally STAR is very fast and it can be easily used in environments where computational power is limited. Before the actual mapping of reads we generated a genome index using the latest version of the organism's reference genome in FASTA format. STAR uses indexed genomes in order to perform fast non-linear searches for the MMP on the reference genome. Thus, how the genome is generated affects the performance of the following steps. Namely a lot of frequent indices improve speed but need more memory, while less, more sparse indices create a trade off between memory consumption and computational speed. We created the genome index using the "sparse indices" option in order to reduce memory consumption. This makes a reasonable trade-off to algorithm's speed but in most cases it does not significantly affect execution time. This mapping step is the most computationally expensive step of this data analysis pipeline but it has to be executed once and its output is used for all subsequent data analysis. In order to keep things consistent we always used the same resource for the reference genome and for fetching the sequences of the predicted LncRNAs. In every case this was the latest version of the ENSEMBL genome in FASTA format.

We then created BAM files sorted by the genomic coordinate and tab separated text files holding the information about splicing junctions. A BAM files is the compressed, binary version of the Sequence/Alignment Map (SAM) file. SAM files, which are flat text files, are the main file format for storing large nucleotide alignments. BAM files are the standard file type for mapped reads to a reference genome (i.e. nucleotide alignments) in binary format. We run the aligner with default parameters and we did not filter out any splicing junctions in that step, as we did this in downstream analysis.

Subsequently, we merged together technical replicates. As library preparation and sequencing lane effects can introduce significant variance in RNA-seq experiments, sequencing of libraries from different biological conditions is usually multiplexed in a sufficient number of sequencing lanes in order to obtain the desirable read depth and marginalize lane-specific batch effects.

Selecting reads according to overlapping genomic features

As the ultimate aim is to go from the representation of discrete RNA-seq reads to estimated gene models for LncRNAs, the first step, after preprocessing data as presented above, was to select only the reads that can be of use. Intuitively, we discarded all reads that overlapped with known protein coding gene models, reads overlapping possible untranslated regions (UTRs) that flank gene models or belonging to proximal enhancers and promoters and reads that overlapped genomic regions that have a very high probability to code for a protein even if they are not annotated in the current organism.

To do this we used the latest genomic annotations, downloaded programmatically using R from the servers of the University of California, Santa Cruz (UCSC) (Meyer et al., 2012) and ENSEMBL/GENCODE (Yates et al., 2016). Using the GenomicFeatures (Lawrence et al., 2013) and biomaRt (Durinck et al., 2009) Bioconductor (Gentleman et al., 2004) packages we fetched the following annotation sets from UCSC: Reference Sequences (RefSeq) genes (Pruitt et al., 2014) annotation compiled by the

National Center for Biotechnology Information (NCBI), ENSEMBL genes annotation compiled by the ENSEMBL/GENCODE project (Yates et al., 2016) and the XenoRefSeq annotation, which includes all known annotated gene model sequences from other organisms which can be accurately aligned to the genome of the organism of the experiment. XenoRefSeq table has been compiled by NCBI and it consists only of alignments made with BLAT (Kent, 2002) with more than 15% of the sequence aligned and more than 40 bases of non-repetitive unmasked DNA. Furthermore only the top 1% of alignments based on identity level and at least 35% base identity of the query were selected. From these annotations we then selected and filtered out any models describing known and predicted long non-coding RNAs using the respective prefixes “NR” and “XR” for RefSeq annotations and models that conform to the description pattern of “linc rna” or “antisense rna” for ENSEMBL annotations. When dealing with poorly annotated genomes, such as rat, we extended these gene models by 2000bp from each side in order to avoid un-annotated untranslated regions (UTRs), proximal promoters and enhancers or not yet annotated exons belonging to those genes. Finally using functions from the GenomicRanges package (Aboyoun et al., 2013) we calculated the intergenic areas, i.e. gaps outside the exons of these annotations, and we selected only the subset of RNA-seq reads that are falling completely within those gaps outside of the exons of the known protein coding gene models.

Identify expressed regions outside known gene models

We then used the subset of reads selected above in order to produce continuously expressed regions. We call these regions islands of expression. Ideally they should be a close approximation of exons, but due to the way of mapping reads to the genome which involves some uncertainty, the limited read length and the potentially wrong base calling they could be significantly different. RNA-seq and essentially all next generation sequencing methods produce certain coverage (the average number of reads that covers all bases of a genomic interval normalised by the length of the region) peaks (figure 3), these peaks usually offer a close approximation of

exons but there is also some coverage on the intronic areas and as a consequence it is almost impossible to find the exact transcription start and end sites from RNA-seq alone (Ulitsky and Bartel, 2013). Essentially, in every RNA-seq experiment there is a number of reads that will always map in un-transcribed regions, either intronic or intergenic, but there is a certain coverage threshold that could effectively separate those artefacts from actually transcribed regions. We have developed a specific function for performing this step.

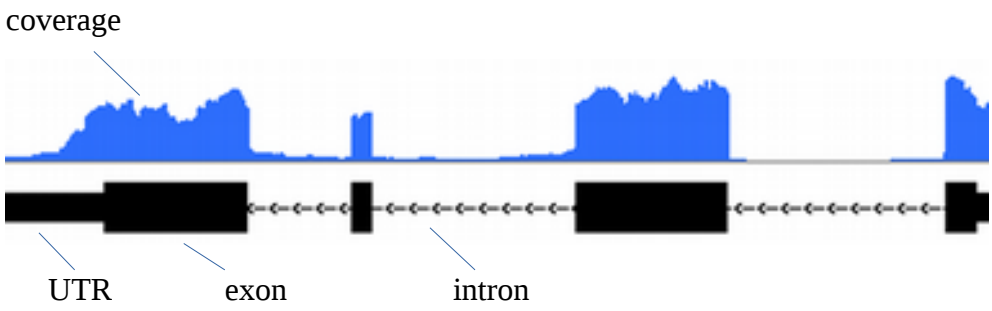


Figure 3: Coverage graph of RNA-seq across an annotated gene model. Coverage peaks are observed above exons (thick black rectangular blocks), but there are also small coverage "leaks" in introns (thin black line). Also the UTR of the gene model (thick black lines at the gene's edges) is being covered by reads. Note that coverage does not stop immediately but rather gradually.

Subsequently, we parsed BAM files containing mapped reads of an RNA-seq experiment into smaller chunks in order to reduce memory usage. In order to fully exploit the advantages of the paired-end protocol, we parsed these aligned reads as pairs of reads using the readGAlignmentPairs function from the GenomicAlignments package (Lawrence et al., 2013). In the context of paired-end reads we do not refer to individual reads but rather to mated pairs of reads. We then subset the reads that exclusively overlap regions outside known gene models as described above.

The next step of the function involves the transformation of these aligned reads into easily manipulated data types. We stored them in a GrangesList (Aboyoun et al., 2013), which is an indexed list of genomic intervals inheriting all features from the R class data.table (Dowle et al.,

2015), an indexed table allowing non-linear search and efficient memory usage for sparse matrices. Then, reads mapped to the forward and reverse strand were separated and the coverage for each strand was calculated.

Coverage is calculated by the formula *(read counts x read length / library size)* and naturally represents the average number of reads that overlap every base of a genomic interval. We then identified continuous regions above a certain coverage threshold that indicates actual transcription activity. To do this we eventually transformed the list of genomic ranges into a list of integers using the “run length encoding” (rle). The “rle” function encodes the array of the raw number of reads overlapping an interval into their respective run length, where instead of the raw number of overlapping reads we used the length of the longest run with a certain value. In that way we compressed data and also we were able to slice this “rle” vector with a certain threshold in order to acquire the longest possible region that is covered by a certain number of reads. In our case we would like to find continuous regions, which can be parts/tiles of the non-coding transcriptome, with at least one read overlapping all their nucleotides with no gaps. Thus we used a cut-off threshold of 1 to slice the run length encoded coverage vector. As islands of expression are defined as expressed regions outside known gene models, the incentive was to identify and further process all these expressed regions. Thus we used this deliberately low cut-off is as we would like to identify the largest possible extent of individual islands of expression and produce a close approximation of full length gene models / loci of LncRNAs. We then finally grouped, trimmed and sliced these, possibly elongated, exons by the *de novo* identified splicing junctions.

To sum up we first calculated the coverage using the fraction of reads that did not overlap protein coding genes. That gave us a representation of the fraction of total RNA that has been sequenced in our library and which could be derived from LncRNAs. Then we sliced that transcriptome into regions longer than 100bp, which is the shortest RNA-seq read length that we can use in the pipeline, that had a continuous

coverage of more than 1. Intuitively this represents the fraction of the non-coding transcriptome comprised of intervals that had their bases covered by more than one read on average. We call these intervals islands of expression, see table 1. This cut-off is intentionally very low, as the goal at this step is to identify the whole fraction of the genome which is transcribed outside known gene models. We consider as evidence of transcription the continuous accumulation of more than 1 read (coverage > 1) for every base of a genomic interval of 100bp. Thus we we used a continuous run of at least 100 bases with coverage > 1 and then we further processed these regions into putative models of LncRNAs downstream in the pipeline.

	Islands of expression
Genomic Context	Transcribed areas that do not overlap with gene models (ENSEMBL, RefSeq, XenoRefSeq)
Length	Length > 100bp (1 RNA-seq read mate length)
Expression Threshold	Coverage > 1

Table 1: Attributes of Islands of Expression

```
> bam
GRanges object with 485279 ranges and 0 metadata columns:
           seqnames              ranges strand
              <Rle>           <IRanges>  <Rle>
       [1]    chr10   [3134454, 3134589]      +
       [2]    chr10   [3135740, 3135860]      +
       [3]    chr10   [3223898, 3224007]      +
       [4]    chr10   [3241880, 3242170]      +
       [5]    chr10   [3243413, 3243683]      +
       ...      ...                 ...    ...
  [485275]     chrY [90507947, 90508085]      -
  [485276]     chrY [90775001, 90775116]      -
  [485277]     chrY [90785703, 90785816]      -
  [485278]     chrY [90793394, 90793580]      -
  [485279]     chrY [90825392, 90825528]      -
  -------
  seqinfo: 66 sequences from an unspecified genome; no seqlengths
```

Figure 4: Genomic Ranges in a GRanges object containing the subset of RNA-seq reads / pairs of reads overlapping with non protein-coding regions

We repeated this process for both strands and then we transformed the results again in a GenomicRanges object, which holds genomic intervals that are sufficiently transcribed in at least one of the BAM files of the RNA-seq experiment. These objects are indexed and hold information on genomic

intervals by storing the chromosome, start position, end position, strand and metadata including a name and possibly a coverage score, see figure 4 for a detailed representation of genomic intervals. At the same time the pipeline creates detailed logs including the number of reads or pairs of reads selected and the number of islands of expression, i.e. transcribed genomic intervals, identified for each BAM file, see table 2. Finally, we created a customised annotation of novel transcribed regions by collecting all those regions identified in each BAM file and collapsing the overlapping and book-ended ones into a single region that spans across all of the collapsed features. In this way we incorporated information by combining read-depth from all samples in an RNA-seq experiment. Thus we have generated a customised annotation of continuously transcribed regions outside known gene models.

Log file for Sample51_BALB.c_SHAM_M.bam
Chunk nr: 1 . Number of reads (pairs) overlapping with intergenic regions: 520048
Chunk nr: 1 . Number of reads overlapping with islands of expression coverage > 1 & width > 100: 55603
Integrating chunk nr 1 to bam file Sample51_BALB.c_SHAM_M.bam. Nr of reads overlapping with intergenic regions: 55603

Chunk nr: 2 . Number of reads (pairs) overlapping with intergenic regions: 188923
Chunk nr: 2 . Number of reads overlapping with islands of expression coverage > 1 & width > 100: 14757
Integrating chunk nr 2 to bam file Sample51_BALB.c_SHAM_M.bam. Nr of reads overlapping with intergenic regions: 70360

Chunk nr: 3 . Number of reads (pairs) overlapping with intergenic regions: 582369
Chunk nr: 3 . Number of reads overlapping with islands of expression coverage > 1 & width > 100: 63916
Integrating chunk nr 3 to bam file Sample51_BALB.c_SHAM_M.bam. Nr of reads overlapping with intergenic regions: 134276

Chunk nr: 4 . Number of reads (pairs) overlapping with intergenic regions: 1444413
Chunk nr: 4 . Number of reads overlapping with islands of expression coverage > 1 & width > 100: 57571
Integrating chunk nr 4 to bam file Sample51_BALB.c_SHAM_M.bam. Nr of reads overlapping with intergenic regions: 191847

Chunk nr: 5 . Number of reads (pairs) overlapping with intergenic regions: 496979
Chunk nr: 5 . Number of reads overlapping with islands of expression coverage > 1 & width > 100: 54005
Integrating chunk nr 5 to bam file Sample51_BALB.c_SHAM_M.bam. Nr of reads overlapping with intergenic regions: 245852

Chunk nr: 6 . Number of reads (pairs) overlapping with intergenic regions: 756139
Chunk nr: 6 . Number of reads overlapping with islands of expression coverage > 1 & width > 100: 71220
Integrating chunk nr 6 to bam file Sample51_BALB.c_SHAM_M.bam. Nr of reads overlapping with intergenic regions: 317072

Chunk nr: 7 . Number of reads (pairs) overlapping with intergenic regions: 601027
Chunk nr: 7 . Number of reads overlapping with islands of expression coverage > 1 & width > 100: 59339
Integrating chunk nr 7 to bam file Sample51_BALB.c_SHAM_M.bam. Nr of reads overlapping with intergenic regions: 376411

Chunk nr: 8 . Number of reads (pairs) overlapping with intergenic regions: 779006
Chunk nr: 8 . Number of reads overlapping with islands of expression coverage > 1 & width > 100: 73097
Integrating chunk nr 8 to bam file Sample51_BALB.c_SHAM_M.bam. Nr of reads overlapping with intergenic regions: 449508

Chunk nr: 9 . Number of reads (pairs) overlapping with intergenic regions: 477328
Chunk nr: 9 . Number of reads overlapping with islands of expression coverage > 1 & width > 100: 45811
Integrating chunk nr 9 to bam file Sample51_BALB.c_SHAM_M.bam. Nr of reads overlapping with intergenic regions: 495319

Table 2: Log file showing all the rounds of read selection and island of expression identification. In order for the pipeline to be memory efficient BAM files are processed in chunks of 4500000 pairs of reads. For each chunk of a BAM file the 1st line shows the total number of pairs of reads outside known gene models for this particular chunk, the 2nd line shows the number of reads overlapping continuous islands of expression for this particular chunk and the 3rd line shows the cumulative total number of reads overlapping islands of expression for this BAM file. In this particular example we can see the log file for Sample51_BALB.c_SHAM_M.bam. In order to completely parse this particular file and identify expressed regions outside known gene models we processed 9 chunks. At the end we had 495319 reads that were overlapping continuously transcribed regions with coverage more than 1. These islands of expression (produced by 495319 pairs of reads) comprised our customised annotation of putative LncRNAs for this particular BAM file.

Reconstruct genes of putative LncRNAs

These islands of expression would include all previously unknown exons of novel LncRNAs but we still need to trim and group them together into transcripts and filter out those that are below the minimum length threshold of 200bp, see figure 5. Additionally, as RNA-seq, due to its nature as a collection of short discrete reads, mapped to a genome using a process which involves some uncertainty, cannot produce accurate start and end sites of transcription, we need to trim these islands of expression according to a scaffold created by the identified splicing junctions. Essentially we used two sources of information, one is the *de novo* identified splicing junctions and the other is the islands of expression outside known protein coding gene models, which have their nucleotides covered by at least one read for their full length. By combining together these two sources of information we acquired a prediction of unknown gene models. If these gene models are more than 200bp in length and with no coding potential, they are putative LncRNAs. In order to reconstruct these gene models we filtered out spurious splicing junctions and islands of expression which were more probable to be false positives.

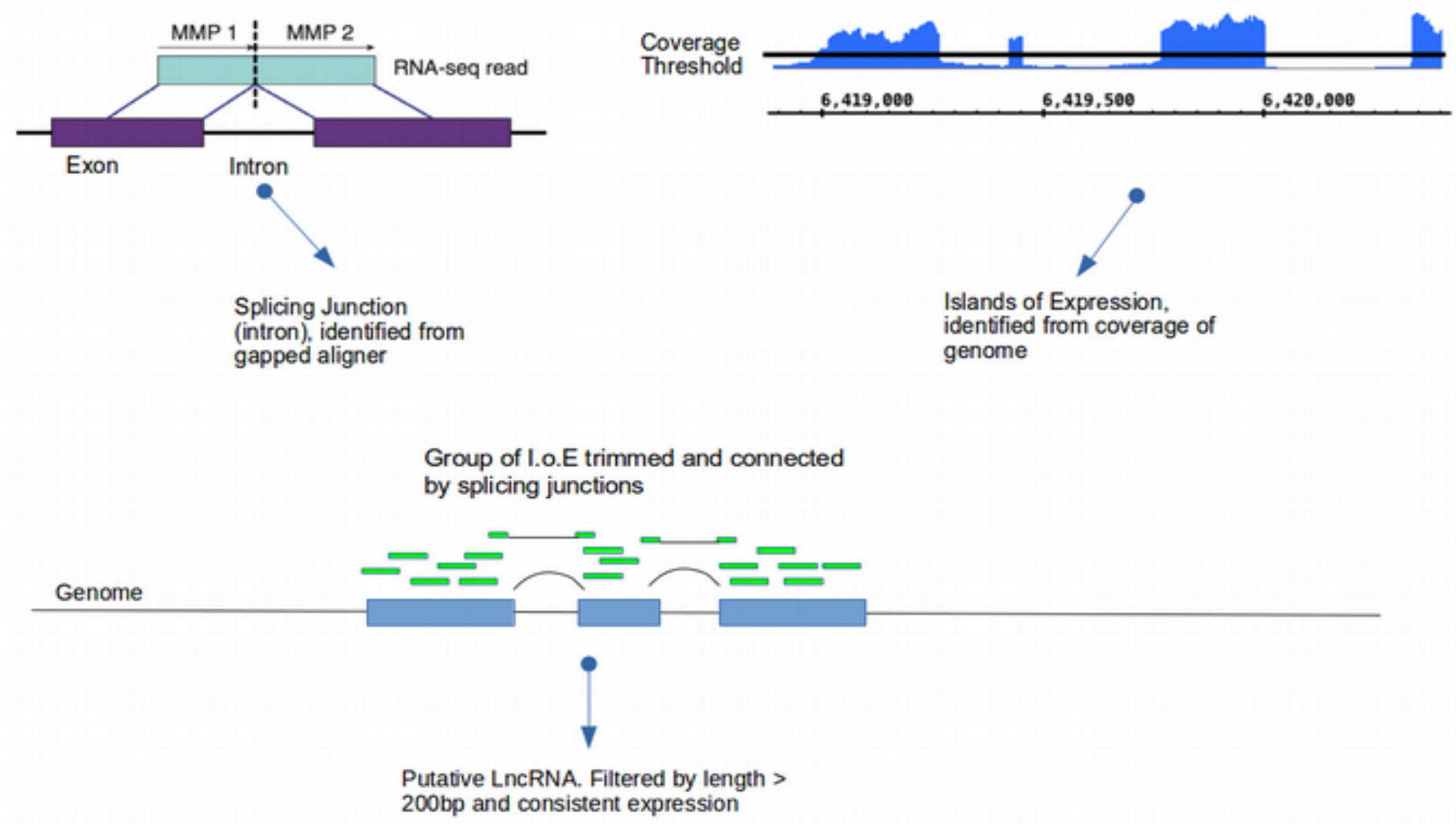


Figure 5: Islands of Expression are being grouped together and trimmed by Splicing Junctions into putative LncRNAs.

We first selected the subset of novel splicing junctions which were confidently identified. Splicing junctions were stored in tab separated text files that indicate the genomic interval of the respective intron as well as the type of the donor and acceptor site, the read overhangs and read counts. More specifically we set the maximum overhang of the read mates of pair of reads on the donor or acceptor site to be (*read mate length) – 1*. The maximum overhang of RNA-seq reads to the donor or acceptor site dictates how many bases of a read are allowed to overlap either the donor or acceptor site when a read is spliced and a splicing junction is identified. In the case of paired-end reads we used as the maximum overhang the length of one read mate, which in our case is 100b, minus 1. This means that at maximum a read could be spliced and mapped using 99b on one side. Thus splicing junctions were identified only when the original mates were spliced and the sequence of one of the read-mates was mapped on both sites of the splicing junction. Splicing of the fragment (an not the read-mate itself) sequenced from both ends with paired-end reads would not identify a splicing junction if the original mates were not spliced. Moreover during the mapping step we used the default minimum allowed overhang of 3b. This means that overhangs of 1 and 2 bases were prohibited for the identification of splicing junctions.

We also imposed a cut-off threshold of more than 2 reads overlapping a junction to consider it a valid splicing event. This threshold produces the lowest ratio of mapper-specific junctions (Dobin et al., 2013) , thus splicing junctions accumulating reads above this threshold are consistently identified by different aligners and are less likely to be false positives.

We parsed in R, collapsed and sorted all splicing junctions detected. Subsequently we developed a function that filters out spurious splicing junctions and groups together islands of expression transforming them into putative exons by trimming the edges according to the identified splicing activity. Grouping of islands of expressions and splicing junctions was done as shown in figure 6.

To do this we exploited a feature that relates to the way splicing junctions are identified. As reads are mapped using the MMP concept and we used the same reads to identify islands of expression and splicing junctions, by definition the very same reads whose aligned seeds have produced a certain coverage value over a genomic interval, have also identified novel splicing junctions as they were split into seeds in the first place. Some islands of expression would overlap the same junction, some junctions would overlap more than one island of expression and vice versa. In that way we defined co-overlapping sets of junctions and islands of expression that have been connected by the same RNA-seq reads, see figure 6. After this first grouping, we got a set of splicing junctions and islands of expression stored in the respective vectors. Some of them might overlap each other, essentially describing parts of the same genomic interval. Thus, we collapsed overlapping islands of expression and we selected from the overlapping splicing junctions the regions which are more probable to represent actual introns. We call the latter regions consensus junctions.

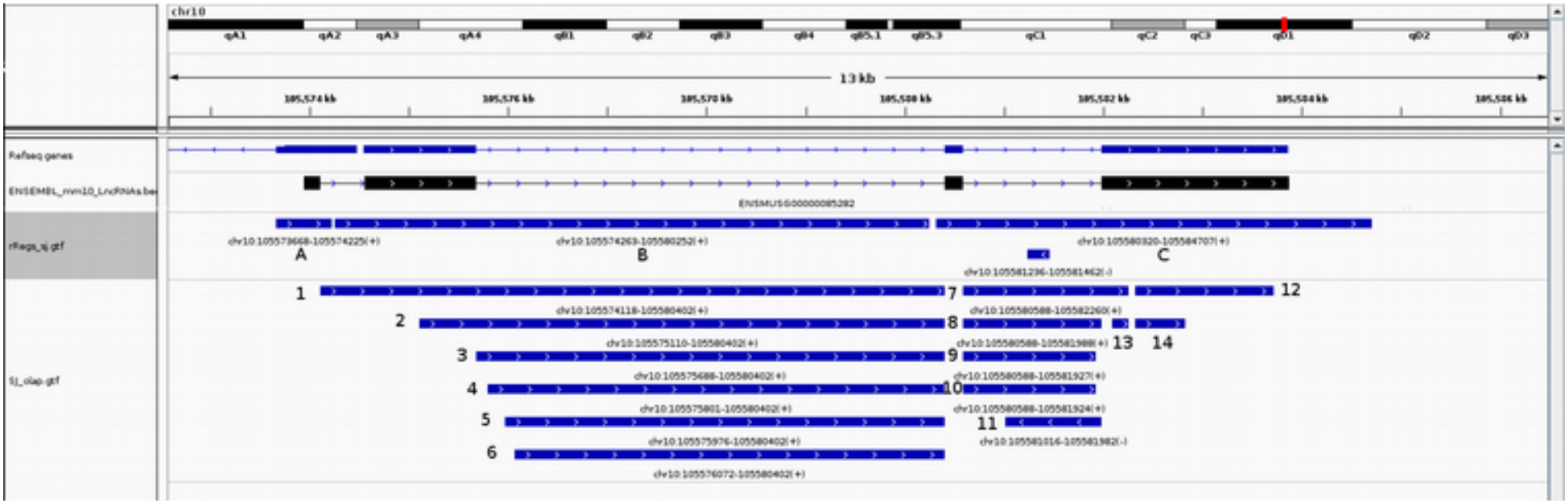


Figure 6: 3 islands of expression (I.oe), A, B and C and 14 splicing junctions (SJ) define a set of co-overlapping features. As a matter of fact I.o.e A overlaps SJ 1, I.o.e B overlaps SJ 1,2,3,4,5,6 etc. We can create three vectors:

A → SJ 1

B → SJ ,12,3,4,5,6

C → SJ 1,2,3,4,5,6,7,8,9,10,11,12,13,14

These vectors have common features so I.o.e A, B and C can be grouped together alongside with SJ 1-14.

ENSMUG00000085282 is an annotated ENSEMBL LncRNA, the 3 continuous islands of expression A, B and C with coverage > 1 And length > 100b gave a good approximation of the maximum transcribed loci. Further downstream processing would trim these elongated transcribed regions according to the set of splicing junctions comprised of SJ1 to SJ14.

To select consensus junctions of high-confidence we calculated the smallest set of genomic intervals that are non overlapping and which together, reproduce the initial regions described in the novel splicing junctions annotation. This is called a disjoint transformation of the genomic intervals and by definition it consists only of non overlapping features, or in other words it splits the ranges into non-overlapping segments keeping track of all the original regions. Then we counted how many of the original splicing junctions overlap with the disjoint ones and thereby we calculated the relative frequency of each discrete segment of a set of splicing junctions. This map of overlaps was used as a “grouping guide”, see figure 7. We note again that *de novo* identified splicing junctions are represented in the form of the genomic intervals of introns.

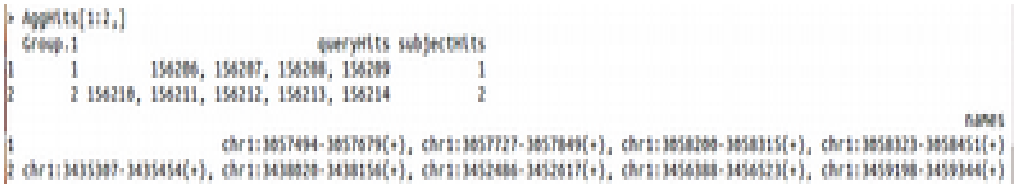


Figure 7: Hits of original splicing junctions (queryHits) on the disjoint splicing junctions (subjectHits). The character vector “names” defines a list of co-overlapping junctions.

We then grouped together all islands of expression that are overlapped by the same set of junctions, using this “grouping guide” as an index. In this way we produced a set of introns and a set of islands of expression that together can be used to reconstruct a gene model of a novel LncRNA. Nevertheless, these junctions do not always consistently describe introns since, some disjoint intron intervals are overlapped by more novel splicing junctions than others. In other words, some of these introns are more probable to be actual introns and thus are identified more frequently than others in the same set. Therefore we looked for a set of consensus introns in which we had higher confidence. Moreover, as we needed to generate an annotation of putative LncRNAs at the gene level and then proceed to differential expression analysis, we had to identify all possible exons of the novel LncRNA gene. Disjoint introns with the maximum

number of overlaps for a specific set of islands of expression and introns / splicing junctions should represent the most dominant transcript, but there might be additional junctions and exons belonging to different transcripts for the same loci and also some junctions could accumulate slightly less reads due to random effects. As we would like to reconstruct the most complete gene models we calculated the top counts and the standard deviation of counts for each set of the disjoint intervals. We always selected the introns with the most counts and also the introns that have counts equal or more than *round(top(counts) – sd(counts))*. These consensus introns represent the high-confidence set of introns for a specific LncRNA model and we used them to trim the islands of expression accordingly. To do so we subtracted the genomic intervals of these consensus introns from the genomic intervals of the grouped islands of expression. That gave us a putative gene model of a novel LncRNA, see figure 8 for an illustration of the consensus introns / splicing junctions and gene model reconstruction.

In figure 8, we present an example of how our pipeline identified a known LncRNA using RNA-sequencing data. The consensus introns / junctions gave an accurate representation of the actual splicing events and the islands of expression gave an almost accurate representation of where transcription starts and ends. The difference between the tracks of islands of expression and the respective consensus introns gave as a very accurate representation of this LncRNA's gene model. We should note here that the differences between the gene models for the this LncRNA amongst the two major annotation projects, RefSeq and ENSEMBL/GENCODE, are more significant than the difference between our prediction and the ENSEMBL annotation.

The pipeline stores and exports these putative LncRNAs in the Gene Transfer Format (GTF), which stores the genomic coordinates of each exon, their order in the transcript and the genomic intervals of the whole RNA of the LncRNA together with a LncRNA name in the form of chrX:start-end(strand). Given the reconstructed transcripts produced by the above process we selected only those which had more than 200bp exonic length

and we fetched their sequence from ENSEMBL genome assemblies in FASTA format.

Coding potential of the complete models of putative LncRNAs was calculated using the Coding Potential Calculator (CPC) (Kong et al., 2007), which is essentially a Support Vector Machine (SVM) classifier which classifies transcripts into potentially protein coding or non-coding. CPC uses two sources of information for classifying transcripts according to their coding potential. One is homology of the translated sequences with known proteins. To do this all the translated FASTA sequences were blasted using a locally installed BLAST+ (Camacho et al., 2009) suite against a local copy of the UniRef90 protein database (Suzek et al., 2007). The more hits the more likely it is for a transcript to be protein-coding. Then CPC parses the BLAST+ output and together with sequence linguistic features it trains a SVM classifier and assigns a score to every transcript. These biologically plausible features include: the quality and coverage of a predicted Open Reading Frame (ORF), the integrity of the ORF that includes whether an ORF begins with a start codon and ends with an in-frame stop codon and the length of ORFs. We filtered out all transcripts with a cpc score > 1. CPC scores between -1 and 1 are considered to be in the grey zone, while scores above 1 are considered definitely coding and scores below -1 definitely non coding. Moreover CPC is more accurate in identifying coding than non-coding transcripts (Kong et al., 2007) and some LncRNAs are known to produce small peptides (Ulitsky and Bartel, 2013). For these reasons, in order no to discard any putative LncRNAs, we only discarded predicted transcripts with a CPC score > 1. Finally, the predicted gene models which were consistently and sufficiently expressed, non-coding and more than 200bp in length comprised a set of putative LncRNAs, see table 3.

	Putative LncRNAs
Annotation	Spliced groups of Islands of Expression and introns / novel SJs
Length	> 200bp
Expression	> 0 reads across at least all samples of one condition
Coding Potential	< 1 CPC score

Table 3: Attributes of putative LncRNAs

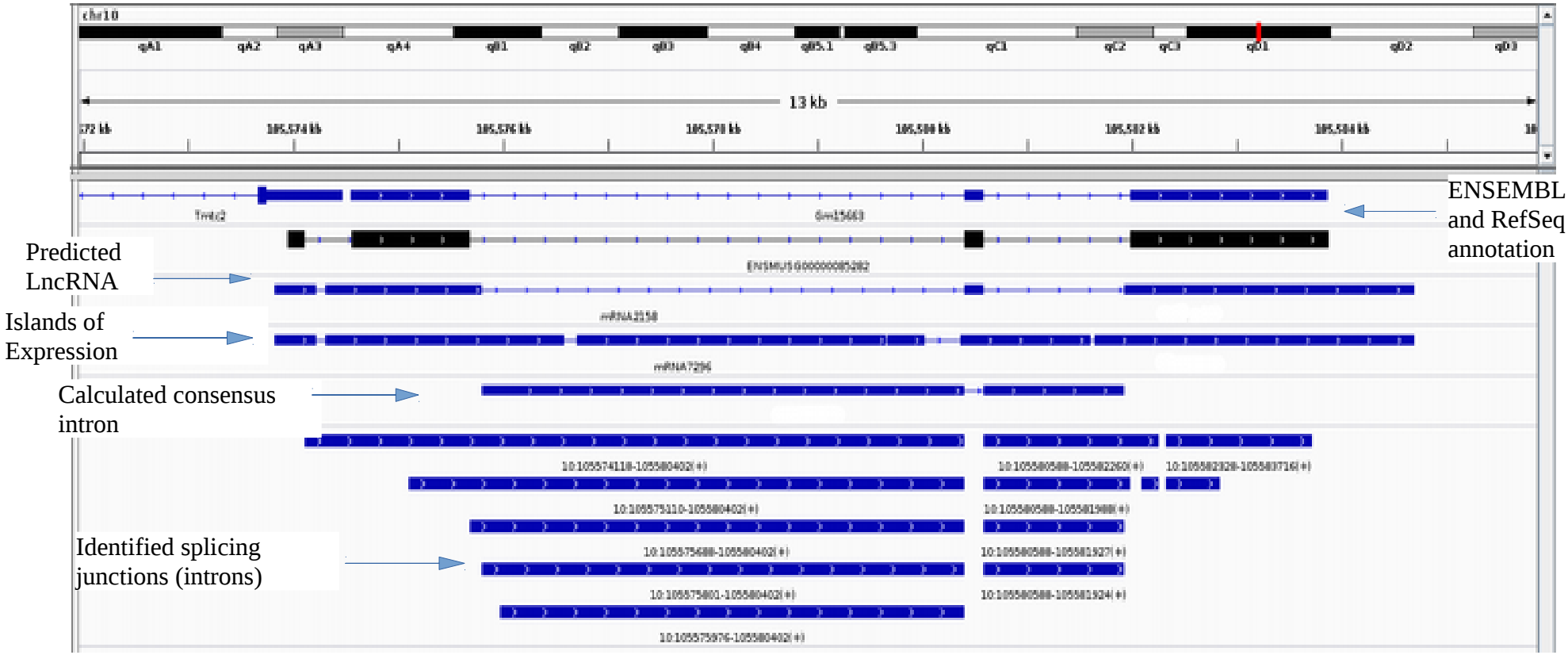


Figure 8: In this figure we present the known mouse LncRNA ENSMUG00000085282 and how our pipeline reconstructed it. Subtracting the consensus intron (calculated from the regions most frequently covered by the identified splicing junctions) from the track showing the de novo identified grouped islands of expression gives us a close approximation of the real gene model. First we calculated the disjoint transformation of introns. Then we counted the overlaps of the original introns / novel splicing junctions on these disjoint regions. This produced the calculated consensus intron. Then we subtracted these introns in which we have higher confidence from the grouped islands of expression. The only significant difference is that the first exon is smaller than in our prediction, but nevertheless the latest RefSeq annotation does not include the first exon at all.

Calculate DE and associate expression profiles of putative LncRNAs and genes

The reconstructed gene models discussed above were exported in the form of the standard Gene Transfer Format (GTF) file, which stores genomic coordinates for the predicted LncRNAs. The output of this pipeline is suitable for further count-based DE analysis. A table of counts for the predicted LncRNAs can be directly calculated as the first step for DE analysis.

Counts were assigned on LncRNAs using the python framework HTSeq (Anders et al., 2015). We counted the reads overlapping exons and summarized them at the gene level of the LncRNA. As a strategy for solving reads assigned to multiple features or to multiple exons of the same LncRNA (i.e. ambiguous reads), we used the IntersectionNotEmpty strategy. According to this strategy a disjoint version of the annotation of LncRNAs was calculated and reads were assigned on the non-overlapping regions. In this way a fraction of the ambiguous reads can be retained and assigned to the feature they mostly overlap. The total number of ambiguous reads and their distribution across samples was also used as a metric to asses both the quality of our annotation of predicted LncRNAs and the quality of individual samples. Although the number of multi-mappers, i.e. reads aligned to more than one genomic region, are a trait of the RNA-seq sample and the genome assembly, the amount of ambiguous reads in the counting step is a trait of the annotation and the counting algorithm. Thus a poor annotation, which does not effectively capture the actual transcriptome's form, would produce much more ambiguous reads. Additionally, if we observe, in an otherwise consistent distribution of ambiguous reads across samples for the same annotation, spikes of ambiguous reads in certain samples and if this finding is consistent with spikes in the number of multi-mappers, then we can hypothesize that those samples suffer from poor sequencing quality, RNA-degradation or genomic contamination and should probably be excluded as outliers. Before any analysis we assessed the

quality of samples by measuring the number of ambiguous reads and the non-unique alignments produced by the multi-mapped reads.

Next, we filtered out transcripts which were not sufficiently or consistently expressed in our experiment. Typically we removed LncRNAs which did not reach counts of more than a threshold for at least all samples of one condition. In this way we removed all predicted loci which showed both very low and highly inconsistent expression with just some expression spikes. Then we used DESeq2 (Love et al., 2014) to calculate differential expression. DESeq2 estimates log fold changes of genes and infers statistical significance of the observed change. The method moderates the dispersion of genes through empirical Bayes shrinkage based on the average expression strength of genes across all samples. The assumption is that genes with similar average expression have similar dispersion. Additionally, it shrinks log fold changes of genes towards zero according to how much information (counts, dispersion, degrees of freedom) is available. In this way, artificially high Log Fold Changes (LFC) occurring due to the dispersion effect on gene models with low reads, are moderated. This is particularly important for predicted LncRNAs, where read counts are very low and very small changes to the total read counts for a condition could lead to very high log fold changes. Then hypothesis testing tests if the coefficient of the fitted Generalised Linear Model for each gene is significantly different from zero. We used the Wald test to calculate adjusted p.values.

Comparing conditions using Generalized Linear Models (GLMs)

In order to compare conditions and identify DE genes between them we need a matrix with some quantification of gene expression, usually a table of counts derived as described in the “Methods” chapter. In general we start from a matrix $\boldsymbol{E_{ij}}$, where rows (***i)*** are the genes/features and columns (***j)*** are the samples. Then the matrix entries are fitted to a specific distribution, usually modelled after the negative binomial distribution. The negative binomial distribution fits the nature of the problem well as it is a discrete probability distribution and it describes the number of successes in a series

of independent Bernouli trials before a certain pre-specified number of failures occurs.

Most methods calculate sample specific constants, which are proportional to the total number of mapped reads for the sample (i.e. the library size for RNA-seq), and use them to normalize gene abundances. Some other methods calculate gene-specific normalization constants by exploiting distinct attributes of the gene like nucleotide content or transcript length.

Then a Generalized Linear Model is used to explain changes in the expression matrix for each gene. We should note here that the model is not linear, but there is a link function that links the linear predictor to the y values. This is the attribute that distinguishes GLMs from linear models.

First counts are transformed into logarithms and then into a model of the form:

$$\begin{aligned} y_1 &= b_0 + b_1 X_{11} + \cdots + b_p X_{1p} + e_1 \\ y_2 &= b_0 + b_1 X_{21} + \cdots + b_p X_{2p} + e_2 \\ y_3 &= b_0 + b_1 X_{31} + \cdots + b_p X_{3p} + e_3 \\ &\vdots \\ y_n &= b_0 + b_1 X_{n1} + \cdots + b_p X_{np} + e_n \end{aligned}$$

The ***y*** response variable represents the measured quantity of a gene's abundance, i.e. counts, for genes ***1*** *to* ***n***. $\boldsymbol{b_o}$, $\boldsymbol{b_1}$, $\boldsymbol{b_2}$ etc are the coefficients which explain changes in variable **y** for condition $\mathbf{X_p}$. The residual term **e** represents the error and is the measured variance of the true value and the predicted value of ***Y*** given the fit of the GLM.

This GLM can be much more elegantly represented in a matrix form

y = Xb + e:

$$\begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \cdots & \cdots & \cdots & X_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \cdots & \cdots & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} b_0 \\ \vdots \\ \vdots \\ \vdots \\ b_p \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ \vdots \\ \vdots \\ e_n \end{bmatrix}$$

As we have a lot of genes this can be written as $y_{ij}=\sum_r x_{jr}b_{ir}$, where x_{jr} are the design matrix elements and b_{ir} the coefficients. In a simple linear model with only two conditions the two coefficients are the mean of the measurements for one group and the difference between the means of the two groups. The design matrix for such a model with formula ***y ~ condition*** would be:

Samples	Intercept	Condition 2
1	1	0
2	1	0
3	1	0
4	1	0
5	1	1
6	1	1
7	1	1
8	1	1

For the example above, if the formula is a model of the mean then the **Intercept** covariate is the mean of all samples. The **Condition 2** covariate is the difference of the estimation of the mean of samples belonging to Condition 2 compared to all samples. In complex models like the one used for this experiment, which are frequently used when we have more than one condition of interest, or if we want to model a nested interaction or batch effects, we can have formulas of the type ***y ~ condition + type + type:condition.*** For example in this imaginary dataset described by the above model we have condition 1 and 2 and type 1 and 2. In such a case we have an additive model of condition plus type, so we can look at variances related to “condition” by controlling variances introduced by “type” and we have also type specific effects of the condition variable using the nested interaction term “type:condition”. A model with the above formula is very useful when we want to compare condition effects across different strains or cell lines and also when we have two interacting

conditions. The above formula can also be more elegantly written ***y ~ condition * type.***

Given such a model and using the R's formulation of GLM for hypothesis testing we can have the following coefficients of the linear model: *intercept*, *type 2*, *condition 2*, *type 2: condition 2*. Then we need to define the reference level for each factor, lets say condition 1 for *condition* and type 1 for *type.* In order to carry out the comparison of "condition 2 vs condition 1" for the reference type only (type 1), we use the estimated coefficient "*condition 2*". In order to compare "type 2 vs type 1" for the reference condition only (condition 1), we use the estimated coefficient "*type 2*". The interaction term "*type 2: condition 2"* describes only the difference between the effect of condition 2 vs condition 1 in type 2, vs the effect of condition 2 vs condition 1 in type 1. Thus this coefficient, i.e. the interaction term, represents the difference of differences.

It is very important to clarify that the interaction term "type 2: condition 2" does not describe the whole response in condition 2 for type 2, but rather how different that is from the effect of condition 2 in the reference type, i.e. type 1. It describes only the difference in coefficients of condition 2 vs condition 1 for type 2, vs condition 2 vs condition 1 for type 1. This is exactly the difference of coefficients between type 2 and type 1 in the comparison of condition 2 vs condition 1. It explains only the proportion of the effect of condition 2 in type 2 that cannot be explained by the general effect of condition 2. Thus this comparison does not involve the comparison of normalised abundances but rather the comparison of fold changes.

We furthermore assessed the expression consistency of LncRNAs using the Cook's distance. Cook's distance is the difference in the coefficient of a linear model explaining the expression of a gene or LncRNA across samples if we remove a sample and refit the model. We plotted the log10 of the Cook's distance for each loci to detect outliers with inconsistent expression and bar-plots of the average Cook's distance for all LncRNAs in a sample to asses quality of samples.

Annotation of predicted LncRNAs

Next, we assessed the ability of the predicted LncRNAs' expression pattern to separate samples in order to asses whether they carry or not any biological signal relevant to the respective experimental conditions using principal components analysis and hierarchical clustering.

Finally, we annotated putative LncRNAs with a unique name using the form chr:start-end(strand). Moreover, we annotated predicted LncRNAs by the gene names that flank it upstream and downstream and also by the gene name of any antisense protein coding gene. Then, after we calculated differential expression of known gene annotations, we used them to identify correlating and anti-correlating expression profiles of protein coding genes and antisense or intergenic LncRNAs.

More specifically we reported all antisense LncRNAs, all LncRNAs which had a known pain gene (Lacroix-Fralish et al., 2007) as their closest neighbour and all LncRNAs which were antisense of a known pain gene. Additionally, we selected all antisense LncRNAs which had an anti-correlated expression profile with the gene on the opposite strand and were both significantly DE (adjusted p.value < 0.05) for the same comparison. We also selected and reported all antisense LncRNAs that were significantly DE on the opposite strand of significantly DE protein coding genes with an opposite log fold change and all significantly DE intergenic LincRNAs proximal to significantly DE protein coding genes These sets of LncRNAs are comprised of predictions that are highly probable of having a functional role in neuropathic pain.

As all LincRNA had a highly correlated expression profile with their most closest protein coding gene and we had no gold standard of known highly correlated genes and LincRNAs, we used a permutation approach in order to select the most highly correlated pairs of LincRNAs and protein coding genes. Pearson correlation was calculated for all samples using the regularised log transformed normalised counts in order to make counts directly comparable. We then utilized a permutation/randomisation. First we generated random numbers from a normal distribution and used them to pair

features of the datasets of LincRNAs and their closest genes. We only considered significant estimations (p.value < 0.05) of the Pearson correlation coefficient regarding the comparison wise error rate. In this way we calculated the correlation between random pairs of protein-coding genes and LincRNAs, we considered these as false positives. We then did the same for the actual pairs of adjacent protein coding genes and LincRNAs, we considered these as true positives. As we wanted to reduce false positives as much as possible, we adjusted the cost of false positives to be 3 times that of false negatives and we adjusted the optimal cut-off threshold of significant correlation so as to discard as much as false positives without losing a lot of true positives. We then assessed the performance of this classifier by the Receiver Operator Characteristic (ROC) curve and the respective Area Under the Curve (AUC) and calculated the optimal cut-off value of the correlation coefficient, figure 9.

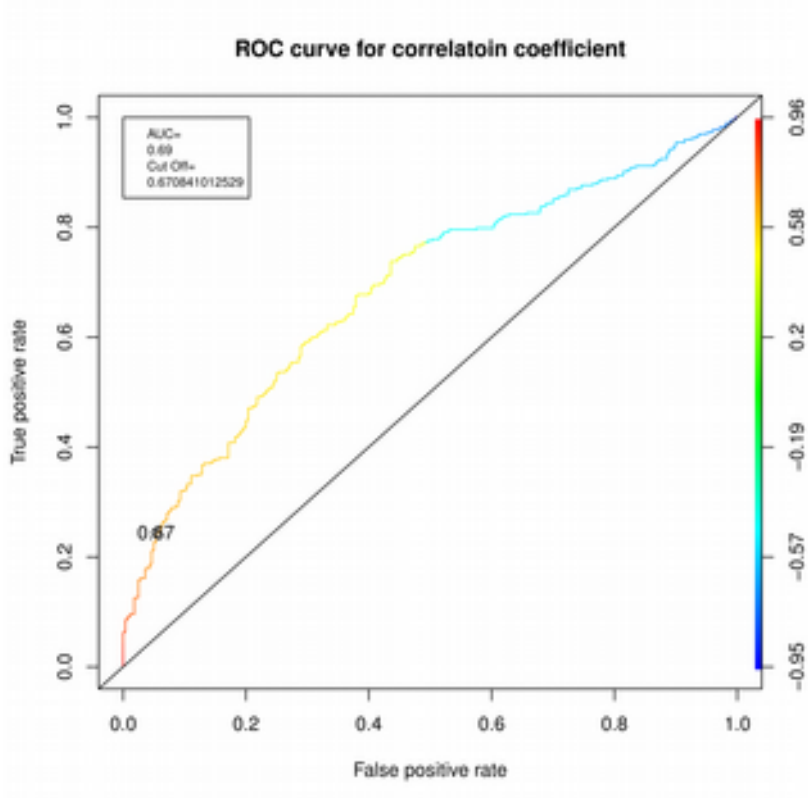


Figure 9: Area under the curve of a classifier built to distinguish between correlation calculated form from random pairing and correlation of actual pairs of proximal genes and LincRNAs. Correlation coefficients above the cut-off were considered significant.

Calculate counts and DE of known genes

Regarding the already annotated and protein coding fraction of the transcriptome, we used the same approach as for LncRNAs. First, we always used the latest annotation of ENSEMBL / GENCODE genes as they are the most comprehensive ones in terms of the inclusion of all possible gene types. They have been also found to be significantly better than RefSeq annotations in terms of the percentage of mapped reads and the number of features with clear evidence support (Seqc/Maqc-Iii Consortium, 2014). Although AceView (Thierry-Mieg and Thierry-Mieg, 2006) outperforms ENSEMBL annotation in human, it is not that widely used and not that convenient to use for DE analysis in automated pipelines. We then assigned counts using HTSeq with the "invert strandedness" option in order to accommodate for the usage of dUTP libraries (see Introduction, section Analysing RNA-Sequencing data) and the "Intersection_nonempty" counting strategy in order to resolve assignment of ambiguous reads and not lose experimental power.

Subsequently, we carried out differential expression analysis using DESeq2. Regarding annotated genes we did not impose any expression filter but we rather relied on independent gene filtering based on Cook's distance as implemented in DESeq2. For performing principal components analysis and for visualization purposes, where transformed counts that remove the variance's dependence on the mean and a zero-centered distribution of counts are highly desirable, we used either the regularized log transformation (rld) or the variance stabilizing transformation (vsd). Both methods borrow information from all samples in the experiment in order to remove the general trend of variance observed on genes with similar mean counts. Also, as we usually dealt with complex designs, we used interaction terms and blocking designs in order to assess group specific responses (see Introduction, section Analysing RNA-sequencing experiments). Finally, we used BiomaRt (Durinck et al., 2009) in order to obtain orthologs between species, to fetch gene symbols and functional annotations and to compile

specific gene lists, like the pain genes list downloaded from the pain genes database.

Functional enrichments

Regarding functional enrichment, we used customised R scripts that rely on functions and classes implemented in the topGO (Alexa and Rahnenfuhrer, 2010) package. We implemented customised functions to adapt the package's functionality to our needs. Our goal was to perform over-representation analysis for gene ontology (Ashburner et al., 2000) terms. Gene ontology (GO) is a project aiming at annotating gene properties with a structured vocabulary divided into three distinct domains. Cellular component, molecular function, which describes gene function at the basic molecular level and biological process, which describes series of molecular events with a certain start and end that are important for the function of organisms, organs, tissues and cells. GO is structured in the form of a directed acyclic graph, where parent terms are the three distinct domains and then, as we go down the tree, we find more specific terms. Thus, terms are organised in a parent-child hierarchy and all child terms can be described, in a more general way, by their parent terms. As the biological process (BP) annotation describes series of biological events of higher order, carried out by organised molecular assemblies, we performed enrichment analysis looking for over-represented GO terms of biological process.

Due to its structure as a directed acyclic graph, GO terms show strong dependencies between them. In addition, some broader terms, higher in the tree structure, which may not capture the actual underlying biological process, can be scored higher than more relevant and specific terms even if we use a hyper-geometric distribution with multiple correction, i.e. when we explicitly take into account the size of the respective GO family.

Several methods have been proposed that take into account the structure of the GO graph. Two of them are implemented in topGO and involve the elimination of genes mapped to significant GO terms from terms that are higher up in the hierarchy and weighting of genes based on the

scores of neighbouring GO terms (Alexa et al., 2006). The elimination method works bottom-up and iteratively removes genes associated with significant GO terms from its ancestors. On the other hand the weight method introduces a scoring system in order to decide if a child better represents interesting genes than its ancestor and thus identifies the most significant local node. We used these algorithms combined with the Fisher Exact test and the Kolmogorov-Smirnof (KS) test as implemented in the topGO package to calculate significance.

In every over-representation / enrichment analysis it is important to define a gene universe and a set of interesting genes according to some parameters. As the universe of genes, i.e. all genes that can be potentially significantly differentially expressed, we usually used all genes than had more than 0 counts in all samples of at least one condition. As the set of interesting genes looking for enrichment we selected either all genes with an adjusted p.value from the DESeq2 Wald-test less than 0.05 or genes selected with a binary function that assigns 0 or 1 to genes found to have an adjusted p.value < 0.05 in certain conditions and strains. When we looked for enrichments in genes that were significantly DE in one species or strain and not in another, or genes significantly DE in a species or strain and also significantly DE in another, we used as the universe of genes all genes that were significantly DE in the reference species or strain. So we looked for enrichments in genes that were commonly, significantly DE in two strains, or in two different species, or genes significantly DE only in one condition, strain or species and not in others. Then we reported the ranking of GO terms by the elimination algorithm with KS test, the weight algorithm with Fisher and KS test and the simple exact Fisher test. We selected and presented enriched terms by their ranking according to the weight algorithm using the Fisher test. We also printed out the GO sub-graph made by the top enriched biological process in order to gain insights in the relationships of the enriched biological process terms. Our analysis does not take into account gene length bias, but is rather an over-representation analysis looking for significantly over-represented GO terms associated with the

significantly DE genes, compared to the gene universe. There are other methods that normalise for gene length biases but DESeq2, which is the method we used for DE analysis, has been proven to have very good performance without normalising for length biases.

In the next two chapters we present results from applying this computational pipeline and we also further discuss its details. We first present results from rats that underwent the SNT pain model and then from mice that underwent the SNI pain model.

Transcriptional changes of protein coding genes and novel LncRNAs in rat's DRG after the SNT pain model

Overview

In order to study the transcriptional changes occurring in well induced pain states, we applied our customised pipeline, presented in chapter Methods, section Reconstruct genes of putative LncRNAs, to Next Generation Sequencing data from rat's Dorsal Root Ganglion (DRG). The pain model we used is the Spinal Nerve Transection (SNT) and for producing controls samples we used sham operated animals. The application of our customised pipeline gives us the advantage of analysing Differential Expression (DE) not only for annotated gene models but also for *ab initio* predicted Long Non-Coding RNAs (LncRNAs). Our hypothesis that certain protein coding genes and novel LncRNAs contribute to neuropathic pain was assessed by identifying novel LncRNAs and neuropathic pain mediators DE between SNT and sham operated rats. We also gained more insights regarding the rat's DRG transcriptome as we identified putative LncRNAs, intergenic (LincRNAs) or antisense of protein coding genes, and associated their expression profiles with that of pain mediators. Surgeries, behavioural tests and tissue extraction were carried out in Steve MacMahon's lab (CARD centre), King's College London, by Ana Antunes-Martins. Library preparation and sequencing was carried out at Oxford Genomics, Oxford. All data analysis was done by me.

Background

The Spinal Nerve Transection pain model

The Spinal Nerve Transection pain model is a model of peripheral neuropathy, where a trauma is surgically introduced in the peripheral nervous system in order to induce reproducible sensory dysfunction, i.e. allodynia, hyperalgesia and spontaneous bursts of pain.

The SNT model involves the tight ligation and transection of the L5 and sometimes L6 spinal nerve branches (figure 1) and produces a well induced and reproducible phenotype with tactile allodynia as its main component (Bennett et al., 2003; Mogil, 2009). In this case only the L5 spinal nerve was transected and the L4 and L6 branches are carefully left intact. All rats received SNT surgeries on their left sciatic nerve branches. Moreover we used sham operated animals, where the spinal nerve has been exposed but not ligated, to generate control samples.

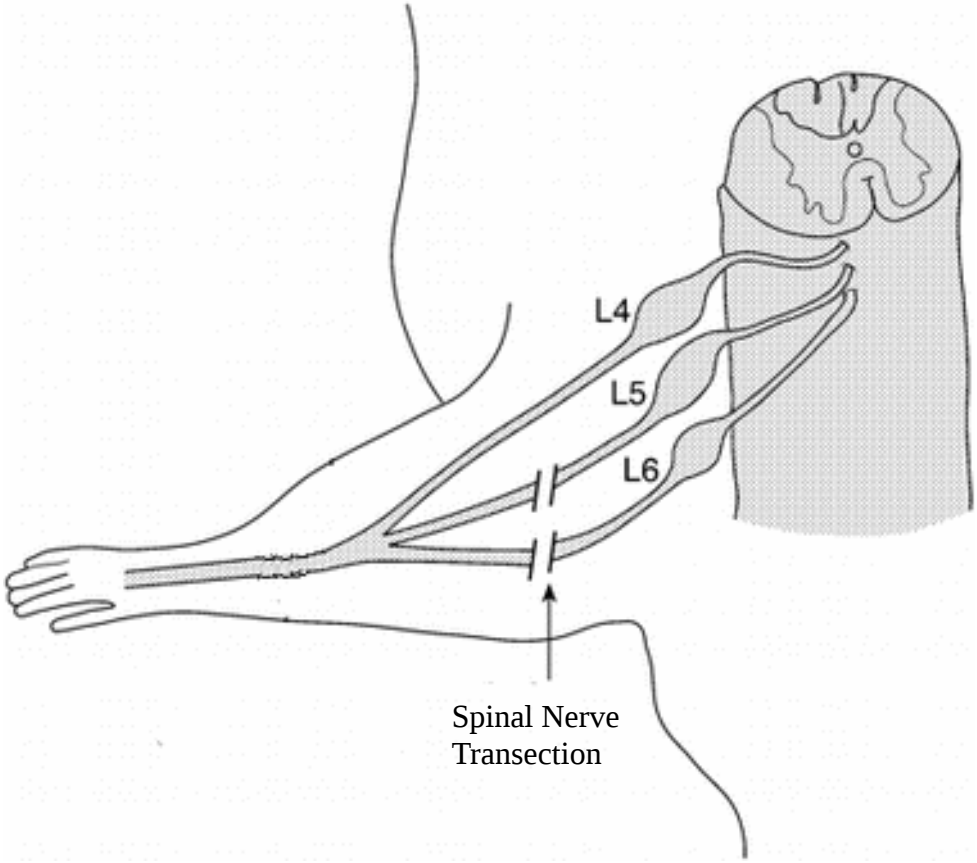


Figure 1: SNT pain model, the L5 and sometimes L6 spinal nerve branches are tightly ligated and transected. The L4 branch remains intact.

RNA-Seq and library preparation

Data was obtained for rats which underwent Spinal Nerve Transection and for rats which only underwent sham surgery and were used as the control group. Tissue from Lumbar 5 (L5) of the Dorsal Root Ganglion (DRG) was harvested and RNA was collected 21 days after the SNT surgery. The timepoint chosen allowed us to assess expression changes due to painful neuropathy, as on Day 21 we assumed that most of the inflammatory response has been resolved.

From each condition four replicates were taken, comprising DRG tissue from 3 pooled animals. Pooling samples of the same condition produces higher quantities, i.e. yield of RNA, for library preparation, and can marginalise unwanted sample specific effects that could introduce an artificially high within samples dispersion. Thus we had 12 replicates for each condition, pooled together into 4 samples for each condition, 8 samples in total. As a consequence of pooling samples together, the 4 replicates we had per condition were not actual biological replicates. Pooling data from multiple animals would eliminate some of the variance making the injury induced molecular response easier to detect and more consistent. All animals were male adult rats, thus we had no sex-effects to control for.

The 8 samples for both conditions were sent for sequencing to Oxford Genomics Center, High Throughput Genomics. Total RNA was provided to the sequencing centre, where the ribodepleted fraction was selected for further sequencing. Ribodepletion excluded the ribosomal RNAs. Subsequently, the ribodepleted RNA was converted to cDNA using the strand-specific (deoxy-UTP strand-marking) dUTP protocol which is the leading protocol for strand-specific synthesis of cDNA. Sequencing was carried out in 4 sequencing lanes with samples multiplexed according to the condition for each lane and with a sequencing depth of more than 50 million reads for each sample.

Aligning RNA-seq reads to genome

Oxford Genomics RNA-sequencing centre produced flat text FastQ sequencing files. Quality metrics were encoded using the Sanger standard and Phred score (Ewing and Green, 1998). This metric assesses the probability that the corresponding base call is wrong. Sequencing was done in four sequencing lanes, producing four technical replicates per sample. In general all these lanes gave high yield, consistent GC content, consistent and expected sequence inserts between the paired-end adapters and high quality base calling (see table 1). Regarding RNA-seq read duplicates, as they are only computationally identified as reads mapped to identical positions, they cannot be separated to PCR amplification artefacts and natural ones. Recent studies (Parekh et al., 2016) suggest that removal of duplicates worsens the false discovery rate of differentially expressed genes and does not improve precision of the analysis. However, in our samples duplicates were very low, approximately 7%.

Lane	% GC	% GC_{mapped}	σ_{pos}(%GC)	insert ± MAD	% exonic	% exon cov'ge	%N	max_{pos} % N	%lowQ	$\%lowQ_{end}$	avgQ
3.1	51.0 ± 10.7	50.7 ± 10.0	3.97	151 ± 42	20.4	89.9	0.0	0.1	0.0	0.0	35.1
3.2	50.9 ± 11.0	50.5 ± 10.6	2.70	150 ± 42	20.5	89.9	0.0	0.0	0.0	0.0	34.5

Lane	% GC	% GC_{mapped}	σ_{pos}(%GC)	insert ± MAD	% exonic	% exon cov'ge	%N	max_{pos} %N	%lowQ	$\%lowQ_{end}$	avgQ
4.1	53.1 ± 11.1	52.8 ± 10.6	3.52	153 ± 42	18.5	89.0	0.0	0.1	0.0	0.0	35.0
4.2	52.9 ± 11.4	52.5 ± 11.0	2.41	153 ± 42	18.5	89.3	0.0	0.0	0.0	0.0	34.5

Lane	% GC	% GC_{mapped}	σ_{pos}(%GC)	insert ± MAD	% exonic	% exon cov'ge	%N	max_{pos} %N	%lowQ	$\%lowQ_{end}$	avgQ
1.1	50.7 ± 10.7	50.6 ± 10.3	4.30	154 ± 42	19.9	89.7	0.0	0.2	0.0	0.0	34.9
1.2	50.7 ± 11.1	50.4 ± 10.9	2.84	152 ± 42	20.0	89.8	0.0	0.1	0.0	0.0	33.3

Lane	% GC	% GC_{mapped}	σ_{pos}(%GC)	insert ± MAD	% exonic	% exon cov'ge	%N	max_{pos} %N	%lowQ	$\%lowQ_{end}$	avgQ
2.1	51.1 ± 10.4	51.0 ± 10.0	3.74	158 ± 46	23.3	90.1	0.0	0.3	0.0	0.0	34.8
2.2	51.3 ± 10.8	51.0 ± 10.5	2.81	156 ± 44	23.4	90.3	0.0	0.0	0.0	0.0	32.8

Table 1: Quality controls for all 4 sequencing lanes.

To produce binary files of RNA-seq reads mapped to the genome (BAM files) used for downstream analysis we used the STAR aligner (Dobin et al., 2013). We mapped all FastQ files to the most comprehensive rat genome assembly (rn5) downloaded from the ENSEMBL genome

browser. Reads were aligned for each sequencing lane in parallel and then the sorted BAM files were merged with their respective technical replicates of the same sample, in order to produce 8 merged, sorted and indexed BAM files. We merged BAM files generated from the same sample in order to acquire the best possible read coverage for all genomic loci, including loci that are transcribed into lowly expressed LncRNAs

As the rat's genome is not that well annotated, unlike the genome of the mouse or human, there was a larger percentage of unmapped reads. 13.94% was the highest percentage of unmapped reads per sample. In general, we had more than 73% of uniquely mapped reads per sample, with the exception of two samples where the uniquely mapped reads were 68.63% and 67.53%. These mapping percentages are good, but not excellent. However, taking into account multi-mapped reads gave us more than 84.5% of reads aligned to rn5 genome in every sample, a percentage which is considered very good and better than similar studies (Gong et al., 2016).

Consequently, we collapsed technical replicates before counting any reads, in order to achieve the best coverage possible in lowly expressed areas of the genome. Then we proceeded to count reads for known genes using the ENSEMBL genomic features annotation, in the form of gene transfer format (GTF) files, programmatically downloaded directly into R using the biomaRt (Durinck et al., 2009) interface. Counting was done in parallel using HTSeq (Anders et al., 2015) with the *intersection not empty* strategy (see chapter Introduction, section Analysing RNA-seq data). We used the recommended options for HTSeq for paired-end sequencing, sorted BAM files by genomic location with strandedness generated from the dUTP library. We assigned counts on the gene level, grouping together multiple transcripts derived from the same gene.

Experimental Design

As we only had one factor of interest with two conditions, i.e. SNT operated animals and sham operated animals as controls, we used a simple design with an intercept term holding the mean of all samples. Log fold changes were moderated and shrunk towards the intercept term, i.e. when we did not have enough information for a specific gene due to high variance or too low counts, its log fold change was shrunk towards the intercept according to the generally observed log fold changes for genes with similar normalised mean counts (see chapter Methods, section Calculate DE and associate expression profiles of putative LncRNAs and genes).

From each experimental condition four replicates were available, enough to calculate within conditions variation and generate accurate estimations of DE. The generalized linear model we fitted for each gene had one coefficient for *condition*, and according to the GLM notation had the following form: ~ *condition.*

Further quality control

Before analysing DE we assessed how RNA-seq multi-mapped reads and reads that were overlapping more than one feature of the annotation, were distributed across samples. As another step of quality control we examined how consistent was the expression of genes across samples.

In BAM files, RNA-seq reads have a hexadecimal header which holds information regarding the quality of base calling but also regarding the way the reads have been mapped to the genome. Reads can be mapped to the genome in more than one positions, but with an optimal mapping position, or in multiple positions with equal probabilities. The latter reads are called multi-mappers. Moreover, in the counting step, reads can be assigned to only one genomic feature / gene model. If this is not the case and the read cannot be confidently assigned to only one feature then the read

is called ambiguous. Multi-mappers can produce a very high number of non-unique alignments in the counting step.

In the context of analysing putative LncRNAs we examined ambiguous reads for both the ENSEMBL annotation and for the customised annotation generated from our pipeline in the form of a GTF file that holds putative LncRNAs. As seen in figure 2, there is very small fluctuation in the number of non-unique alignments between samples and a very stable and small number of ambiguously counted reads in any of the annotations. The non-unique alignments are not the number of multi-mapped reads, but rather how many times these reads have been aligned to the genome and thus identified in counting step. The very small and consistent number of ambiguous reads indicates that all samples are of good quality for DE analysis and also that the customised annotation of LncRNAs has good structure suitable for downstream analysis.

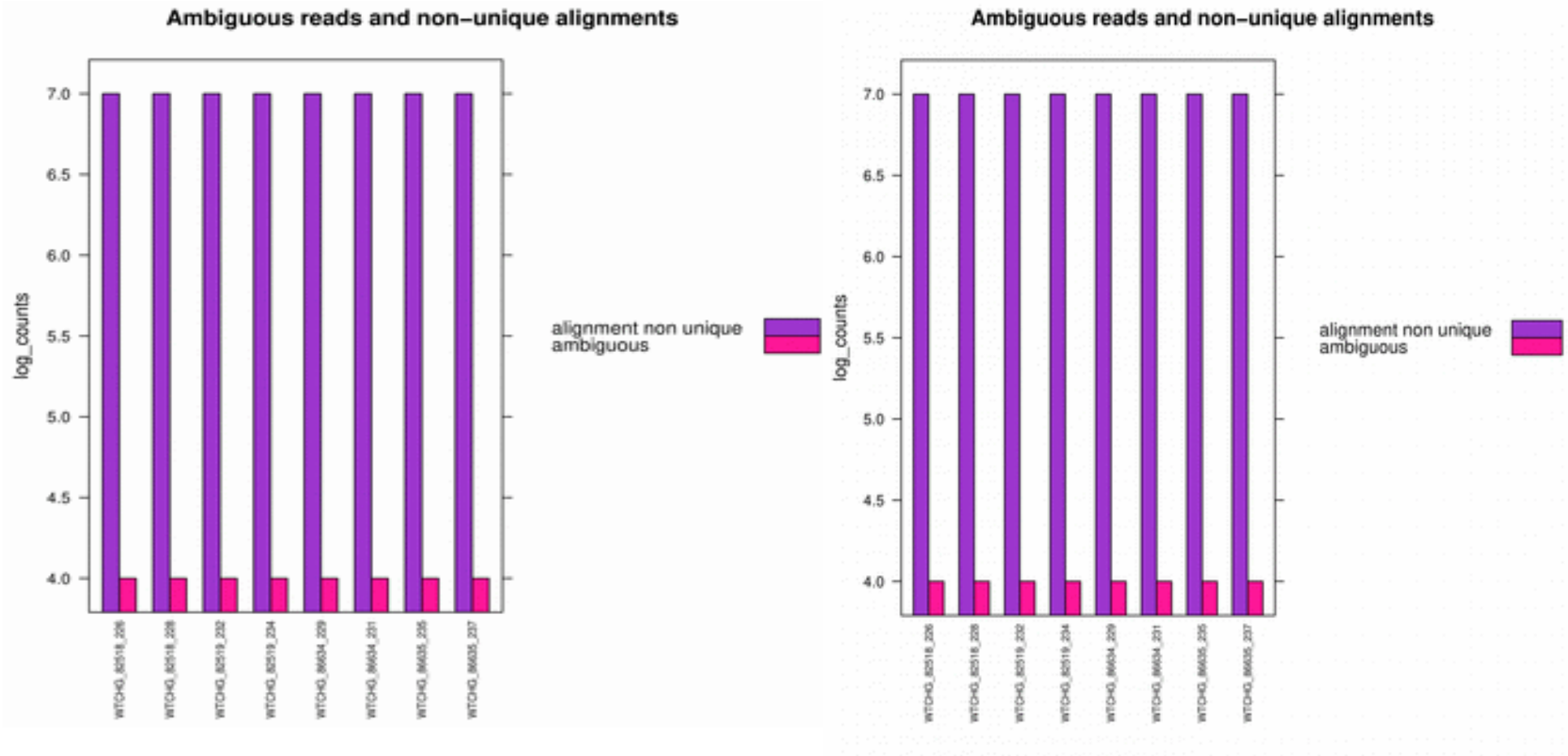


Figure 2: Number of non-unique alignments (violet) and ambiguously counted reads (pink) for ENSEMBL genes (left) and predicted LncRNAs (right). Non-unique alignments arise from multi-mapped reads as they are being counted multiple times. All values have been log2 transformed.

Next we assessed how consistently ENSEMBL genes were expressed across all samples. Genomic contamination or RNA degradation could lead to spurious spikes in gene expression. Thus we calculated the amount that the coefficient of a gene's linear model can change if we remove a sample and refit the model. That is the Cook's Distance and in figure 3 shows the distribution of Cook's distances for all genes in all samples. The median of Cook's distances which is less than 1 (negative logarithm) shows that no particular sample is highly influential to be considered an outlier (Cook and Weisberg, 1982).

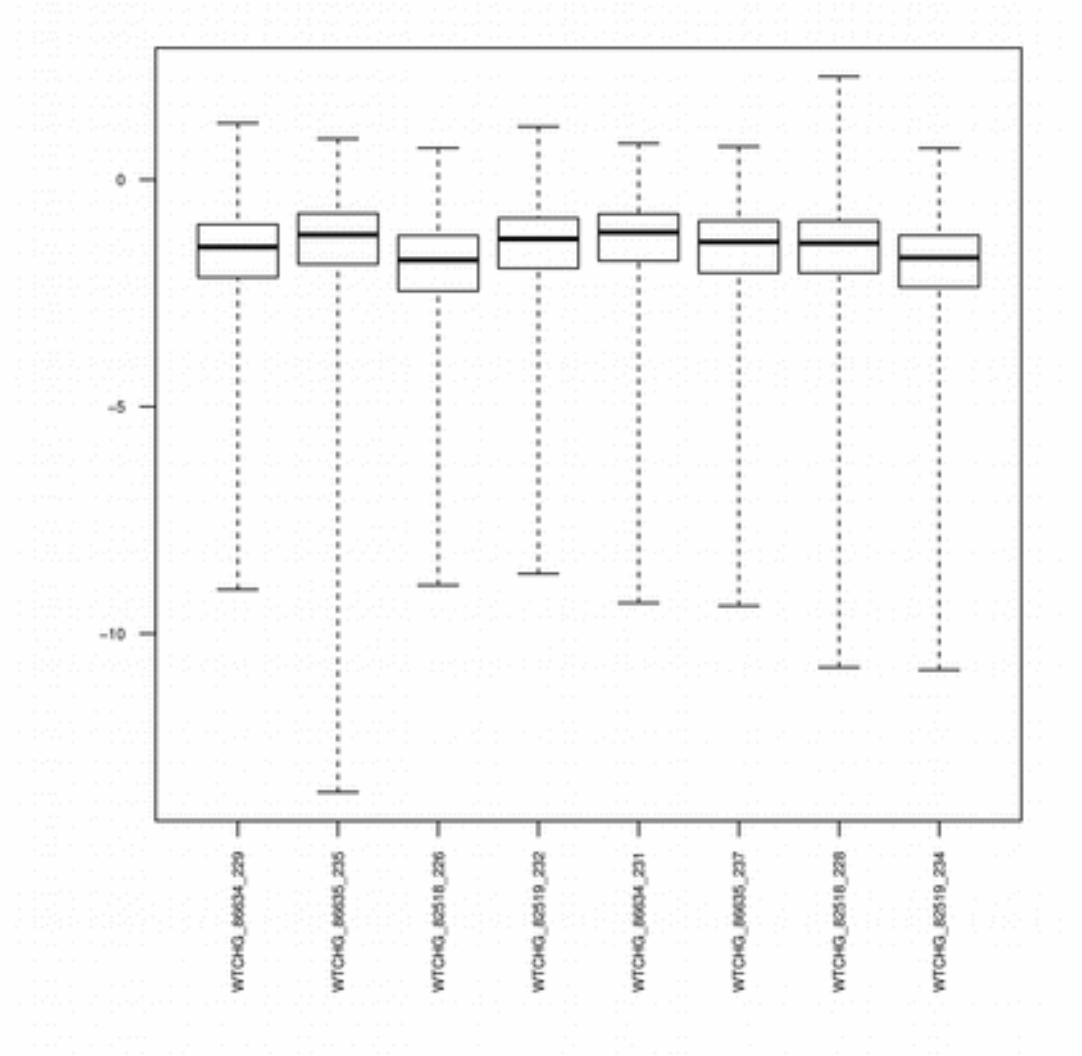


Figure 3: Boxplot of Log10 Cook's distance for all ENSEMBL genes in all samples. From the plot is evident that the majority of Cook's distances is less than 1 and the distribution is consistent between samples.

These results indicated good quality of samples and thus we proceeded to the analysis of DE genes.

Results

Differential Expression analysis of known genes

We used DESeq2 (Love et al., 2014) to analyse count-based RNA-seq data. DESeq2 estimates log fold changes of genes and infers statistical significance of the observed change. For more details regarding differential expression analysis see chapter Introduction, section Analysing RNA-sequencing data and chapter Methods, section Calculate DE and associate expression profiles of putative LncRNAs and genes.

First we normalised and log transformed counts in a way that does not overestimate LFC for genes with low counts, using the regularized logarithm transformation (rld), which shrinks LFC for genes with low counts (Love et al., 2014). The rld transformation also produces a centred mean distribution of counts with equal variances between samples, which makes it ideal for visualisation purposes. Then we clustered samples to examine how gene expression could separate them according to biological condition. All clustering was blind to the experimental design.

All samples were clustered according to their relative euclidean distance. Furthermore principal component analysis was carried out to check if samples were optimally separated by the two first principal components. From the respective figure we concluded all of the samples clustered well with their respective family and different biological conditions could be easily separated using the two first principal components (Figure 4 and 5). The top 500 ENSEMBL genes regarding their variance across all samples were used for these plots.

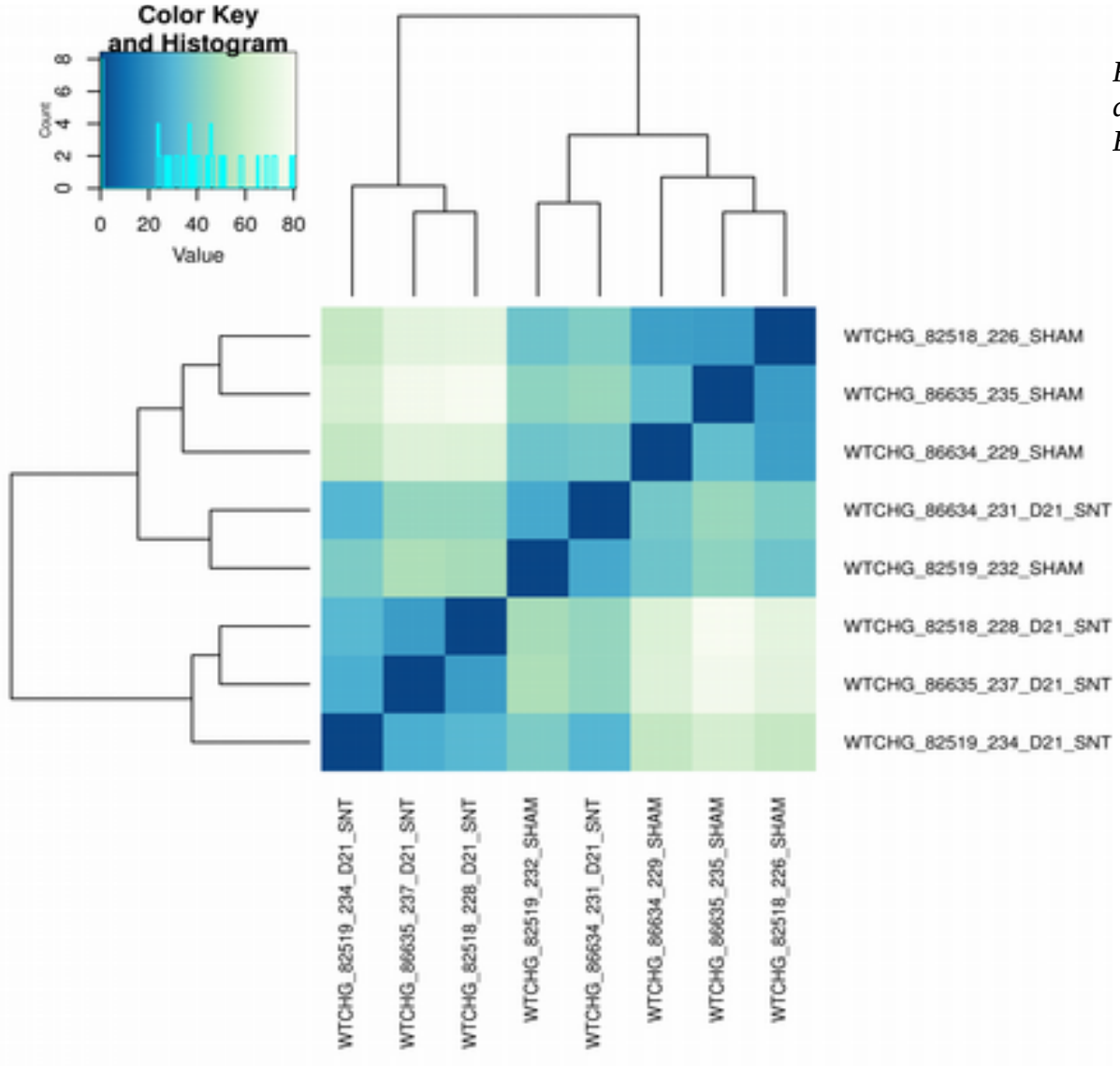


Figure 4: Hierarchical clustering of samples according to regularized log2 counts of ENSEMBL genes.

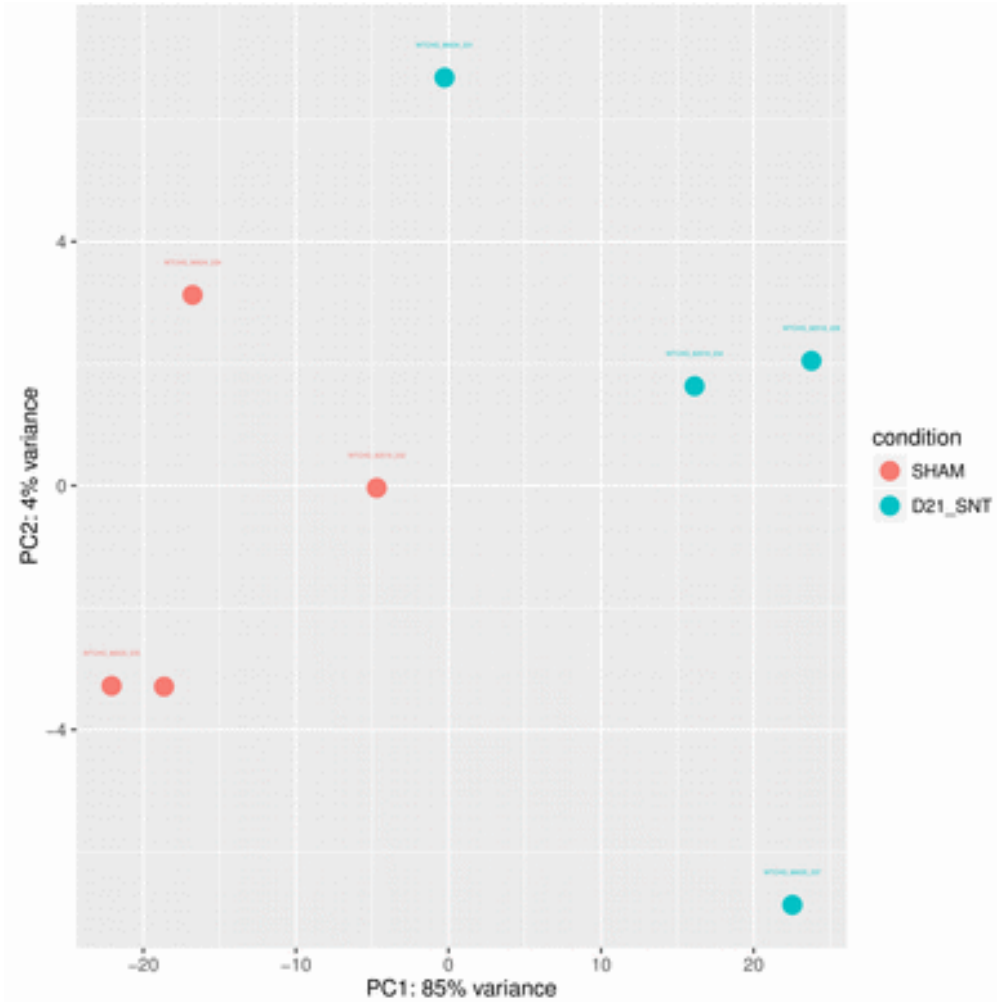


Figure 5: Principal Components analysis of regularized log2 counts of ENSEMBL genes. The top 50 genes contributing to PC1 and PC2 are in Appendix 1.

Significantly DE genes

Using DESeq2 and the Wald test with the Benjamini–Hochberg correction to control false discovery rate and adjust p.values we have assessed the significance of the differential expression of genes. There were in total 14291 genes with nonzero total read counts across conditions. The total number of genes in rat is less than that in human and mouse, this is a direct consequence of the poorer annotation of the rat's genome.

1. ENSEMBL genes significantly differentially expressed (DE) with an adjusted p.value < 0.05 between SNT and Sham operated rats:

   LFC > 0 (up) : 2953, 14%

   LFC < 0 (down) : 3358, 15%

   outliers : 31, 0.14%

The fraction of significantly DE genes identified (29%) implies that major changes are happening in the DRG's transcriptome after the pain model. This is consistent with literature, recently (Wu et al., 2016) found that 1,163 out of of 27,463 (40%) genes were significantly DE in mice DRG after the SNL pain model.

As expected we observed significant transcriptional changes after the well induced pain state of peripheral neuropathy and the amount of genes significantly up-regulated and down-regulated was balanced. Out of the 430 genes functionally validated to be implicated in pain and downloaded from the "pain genes" database (Lacroix-Fralish et al., 2007) 244 passed the independent gene selection of DESeq2 and their DE was calculated. Out of these 244, 115 (47%) had an adjusted p.value < 0.05 and 45 (18.4%) had an adjusted p.value < 0.05 and also a log2 fold change > 1. Thus a significant percentage of these pain genes were found to be significantly DE. The relationship of the log2 fold changes and the p.values for all genes after the pain surgery can be seen in figure 6, pain genes are text label and all genes that had either high fold changes or significant p.values are colour coded.

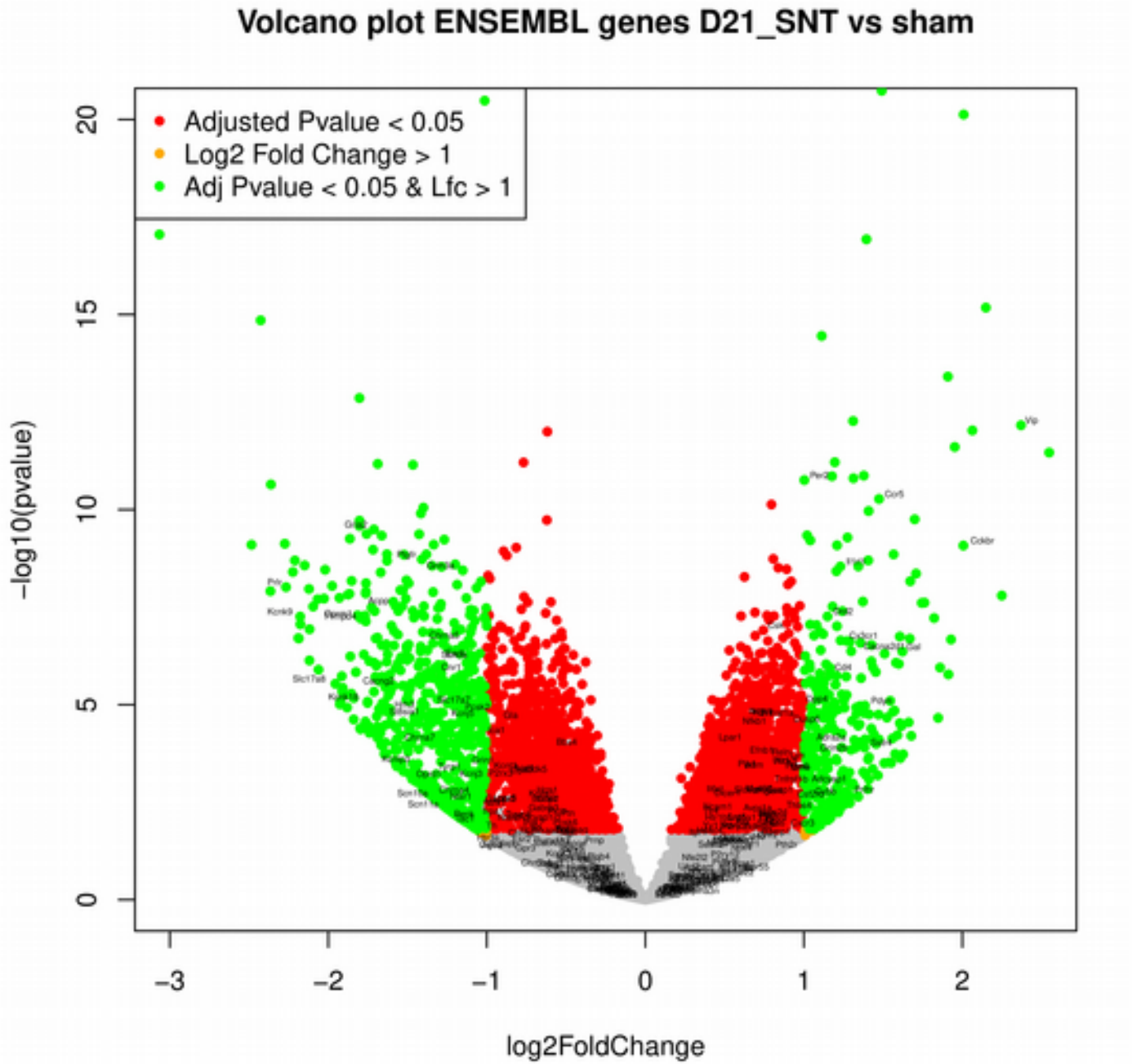


Figure 6: Volcano plot showing the relationship between the log2 fold change of genes and the p.value. Genes that reach very significant p.values < 0.05 and also show high log2 fold change > 1 are colour coded in green. Known pain genes are text labelled on the plot.

Functional enrichment

Based on the above findings that suggested a strong molecular signature of transcriptional changes related to the neuropathic pain phenotype, we proceeded to rigorously assessing the functional enrichment of the set of significantly differentially expressed genes between SNT and sham operated rats. Moreover, as the number of significantly DE genes is large, we can summarize these transcriptional changes in the level of biological processes. To do this we looked for statistically significant over-represented Gene Ontology (GO) terms regarding biological process (Ashburner et al., 2000). In the GO context a biological process is considered any process with a certain start and end where all the distinct steps of the process are accomplished by organised assemblies of molecular functions. We assessed the significance of over-represented GO terms using both count base exact tests, namely the Fisher exact test and some of its variations, and non-parametric hypothesis tests, namely the Kolmogorov–Smirnov (KS) test and its variations (see Methods, section Functional enrichments). We have ranked enriched GO terms by the weighted Fisher test and for completeness we have retained the KS test ranking. To identify the top 20 enriched biological processes we selected the top processes according to the weighted Fisher test that had an exact Fisher test p/value < 0.05. These top 20 enriched GO terms are in table 2 and the GO subgraph generated from the top of them in figure 7.

GO.ID	Term decription	Annotated	Significant	Expected	Rank in weightKS	p.value classicFisher	p.value weightFisher
GO:0071805	potassium ion transmembrane transport	118	84	53.83	1	1.4e-08	7.1e-07
GO:0035725	sodium ion transmembrane transport	89	64	40.6	3	4.1e-07	8.1e-06
GO:0034765	regulation of ion transmembrane transport	265	165	120.89	4	2.6e-08	4.5e-05
GO:0009612	response to mechanical stimulus	198	129	90.33	13	1.8e-08	5.2e-05
GO:0007165	signal transduction	3000	1568	1368.58	6	2.0e-18	6.2e-05
GO:0019228	neuronal action potential	28	23	12.77	2	8.0e-05	8.0e-05
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	499	269	227.64	11	8.6e-05	8.7e-05
GO:0051592	response to calcium ion	95	63	43.34	17	3.6e-05	9.2e-05
GO:0048791	calcium ion-dependent exocytosis of neurotransmitter	21	18	9.58	14	0.00018	0.00018
GO:0090090	negative regulation of canonical Wnt signalling pathway	72	48	32.85	5	0.00024	0.00024
GO:0050770	regulation of axonogenesis	128	78	58.39	36	0.00033	0.00032
GO:0001764	neuron migration	105	63	47.9	30	0.00204	0.00036
GO:0034113	heterotypic cell-cell adhesion	28	22	12.77	32	0.00038	0.00038
GO:0042391	regulation of membrane potential	273	164	124.54	20	8.2e-07	0.00056
GO:0001779	natural killer cell differentiation	13	12	5.93	42	0.00061	0.00061
GO:0048266	behavioral response to pain	16	14	7.3	8	0.00067	0.00067
GO:0071260	cellular response to mechanical stimulus	76	49	34.67	62	0.00069	0.00069
GO:0051965	positive regulation of synapse assembly	51	35	23.27	39	0.00075	0.00075
GO:0070306	lens fiber cell differentiation	19	15	8.67	45	0.00323	0.00085
GO:0007160	cell-matrix adhesion	131	73	59.76	22	0.01240	0.00099

Table 2: The top 20 enriched GO terms ranked according to the KS and fisher exact test. We also present the number of genes annotated under each GO, the number of genes found to be significantly DE and the expected number if there was no enrichment.

We observed a very significant enrichment in GO terms related to ion channels, mainly potassium and sodium. Moreover we had significant enrichments in biological processes related to neuron regeneration and development, such as regulation of axonogenesis and synapse assembly. Furthermore, higher order behavioural terms related to pain phenotype were observed e.g response to mechanical stimulus and behavioural response to pain. Processes related to signalling such as signal transduction, neuronal action potential, calcium ion dependent exocytosis of neurotransmitter are also significantly enriched. This enrichment is consistent with the literature (Lötsch et al., 2013) and also reflects our understanding of the biological processes that are functionally important in neuropathic pain.

As discussed in the chapter Introduction, section Pain at the molecular level, ion channels are some of the main pain mediators and are expressed in nociceptors in order to transmit pain signals and regulate neuronal excitability (Basbaum et al., 2009). Sodium channels transmit information from the periphery to the dorsal horn; potassium channels act as breakers on excitability and are involved in mechanotransduction; Transient receptor potential (TRP) channels are known to activate nociceptors after thermal, chemical and mechanical stimuli. Maladaptive neuronal plasticity and sensation are crucial for maintaining neuropathic pain and the same is true for regeneration of neurons and axons after nerve injury. We can confirm that we observed an extended repertoire of dysregulated genes heavily related to nociception, response to stimuli and pain signal transduction which is consistent with the amount of pain genes dysregulated as seen in the volcano plot of figure 6. In particular terms related to potassium and sodium channels and neuron regeneration were also found to be significantly enriched mainly in BALB/c mice showing high mechanical hypersensitivity after the Spared Nerve Injury pain model and presented in the next chapter. These results are consistent with similar studies in rat DRG after peripheral nerve injury (Gong et al., 2016).

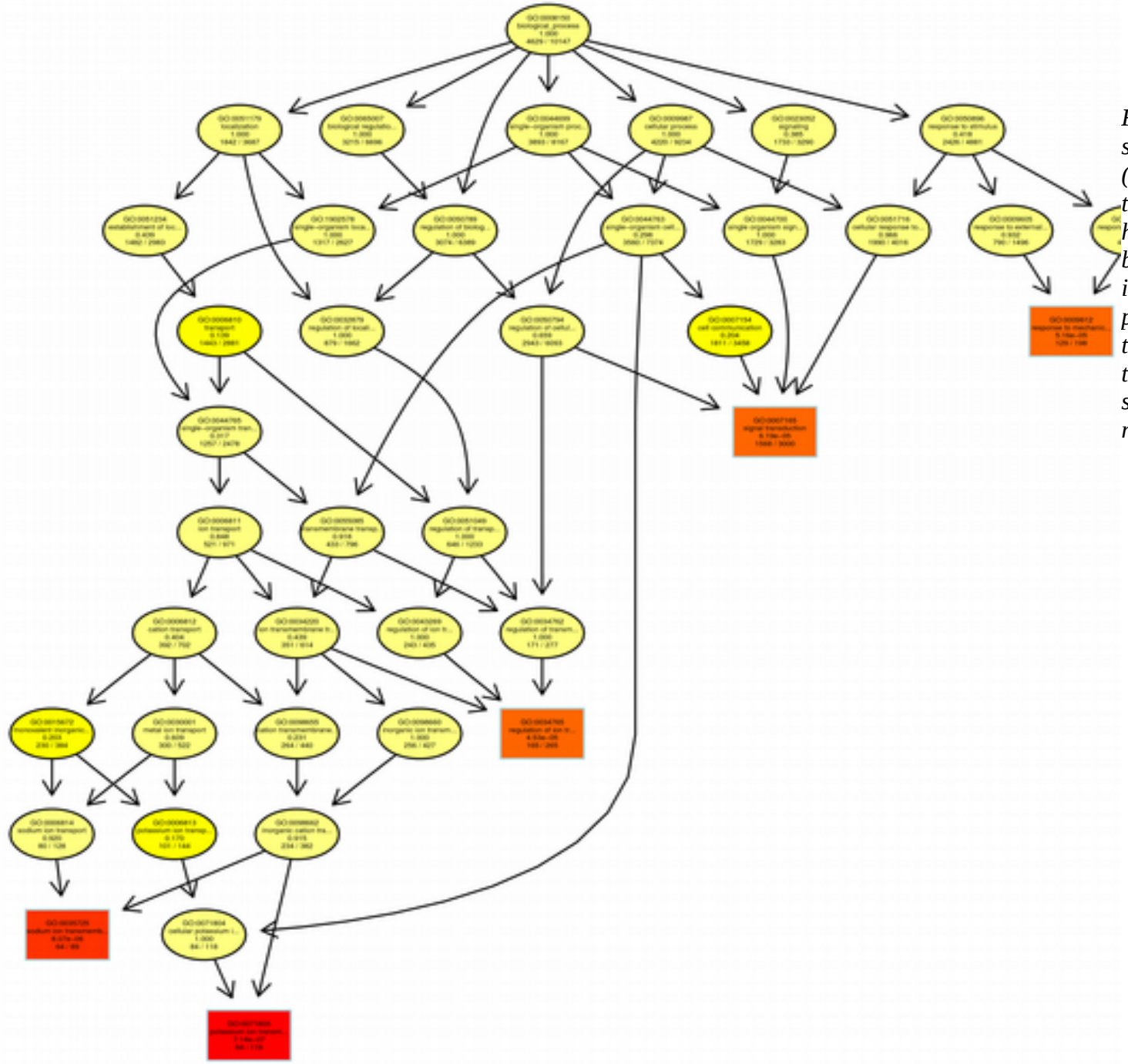


Figure 7: Gene Ontology subgraph leading to the top five (red boxes) highly enriched GO terms. From Left to right the highly significant leaves (red boxes) of the graph are: Sodium ion transmembrane transport, potassium ion transmembrane transport, regulation of ion transmembrane transport, signal transduction, response to mechanical stimulus

Expression patterns of ion channels and pain genes

Guided from the enrichment results and the literature we next decided to specifically analyse ion channels and pain genes and selected all potassium, calcium, sodium, chloride and transient receptor potential channels and pain genes. This was done as these genes are known to be implicated in pain, from our unbiased analysis we found that GO terms associated with these genes were highly enriched and also validated pain genes helped as to identify and confirm a strong molecular response to nerve injury after the pain model. More importantly these genes, which are significantly DE according to our unbiased DE analysis and also functionally important for pain, make a reference point for comparisons between the rat and mouse DRG responses after peripheral nerve injury and for the comparison between the two mouse strains that we will discuss in the next chapter. We plotted their expression patterns using heatmaps and examined how well these specific gene sets could separate samples according to condition.

First, by examining the expression pattern of pain genes derived from the Pain Genes Database (Lacroix-Fralish et al., 2007) (figure 8) we observed that all samples were optimally clustered with their respective families. This finding was consistent with the PCA plot discussed above. Moreover we could identify distinct clusters of pain genes that are co-up-regulated or co-down-regulated after the SNT pain surgery.

Two samples, SHAM_82519_232 and D21_SNT_86634_231 did not have the same transcriptional profile like the other ones. Moreover they did seem more similar to each other than samples from the same conditions respectively. This is evident from figure 8 and we also observed that these samples were closer to each other in the PCA plot (figure 5). As we had no evidence from the quality assessment that there was any RNA quality or RNA abundance issue, or noise due to library preparation or RNA-sequencing, in fact they had excellent metrics like the other samples, we decided not to remove them. We think that reporting of data analysis results should be as complete as possible and inference should be made only based

on evidence gathered from the data itself. Thus we did not want to speculate for a reason why these samples did not follow the expected pattern of transcriptional changes. We should note here, that such an experiment involves a lot of complicated processes, where errors can be accumulated and cause observed data variance.

Regarding potassium channels we observed general down-regulation after the SNT surgery and excellent separation of samples, figure 9. Thus we could confirm the hypothesis that potassium channels, which act as brakes of neuronal excitability, have significantly reduced expression and function in states of peripheral neuropathy leading to neuropathic pain. Moreover transient receptor potential channels had a distinct cluster that is consistently downregulated in SNT animals, contrary to sham, and another cluster which is upregulated and with an expression pattern that can separate samples according to condition in a non-supervised manner. The same was also true for sodium channels, which had a distinct cluster comprised of Scn9a, Scn1b, Scn4b, Scn10a and Scn11a genes encoding voltage-gated sodium channels which were significantly down-regulated after the SNT surgery, figure 10. On the other hand, chloride channels did not show such distinct expression patterns between SNT and sham operated animals, figure 11.

Thus, consistent with the enriched Gene Ontology terms, we could qualitatively observe the significance of ion channels, mainly potassium and sodium, in neuropathic pain and their functional role as pain mediators.

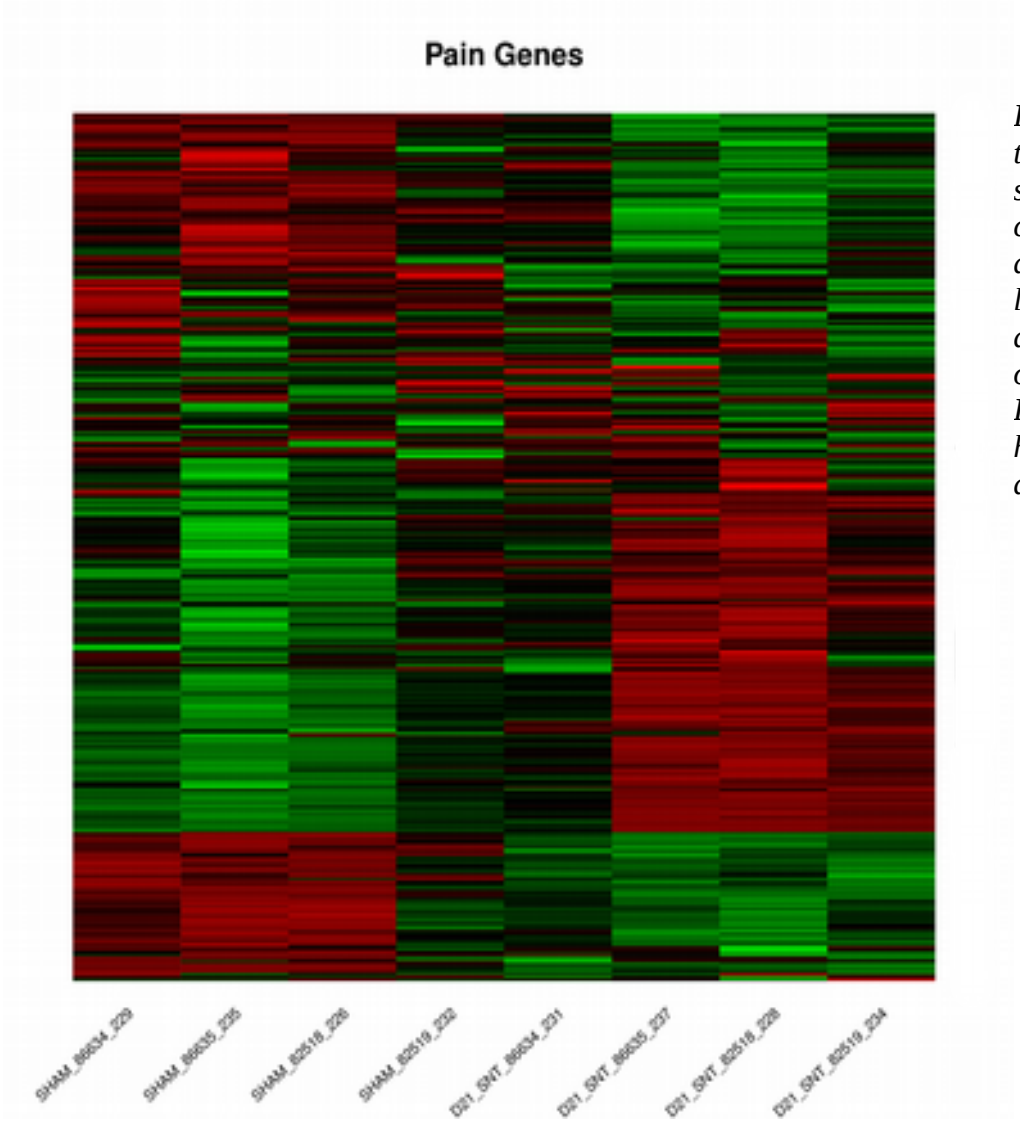


Figure 8: Expression patterns of pain genes based on rld transformed counts. Using only this subset of pain genes samples are almost perfectly classified according to condition and we could also see balanced down-regulation after the SNI surgery (top left) and up-regulation (bottom left). Samples SHAM_82519_232 and D21_SNT_86634_231 did not have such a clear response to nerve injury like the others. We had ruled out any reason that may had to do with RNA or RNA-sequencing quality of these samples. So as we had no evidence from the quality assessment that there was any issue we decided not to remove them.

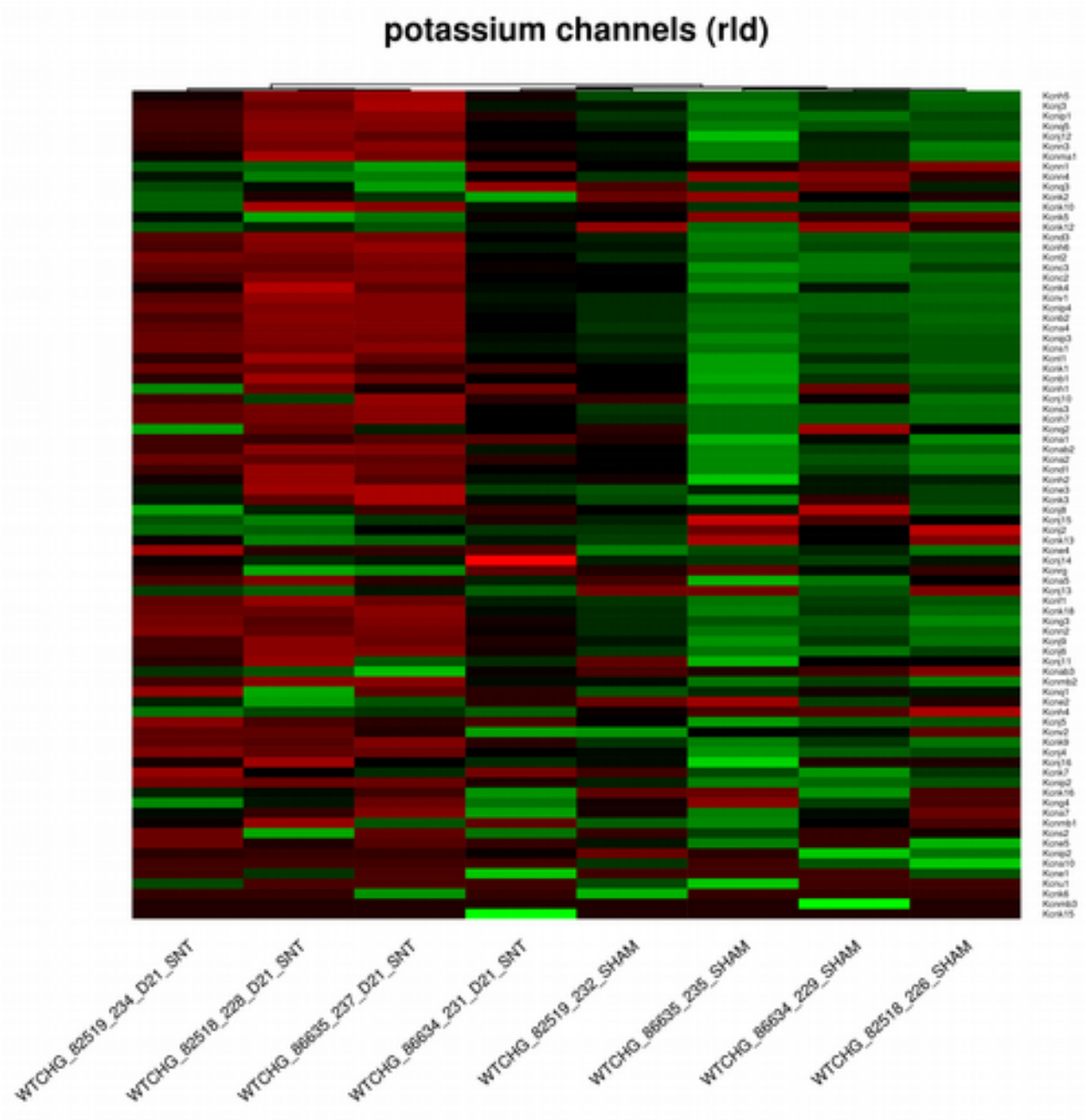


Figure 9: Expression patterns of voltage-gated potassium channels. Most genes encoding voltage gated potassium channels are consistently down-regulated in SNT samples comparing to sham.

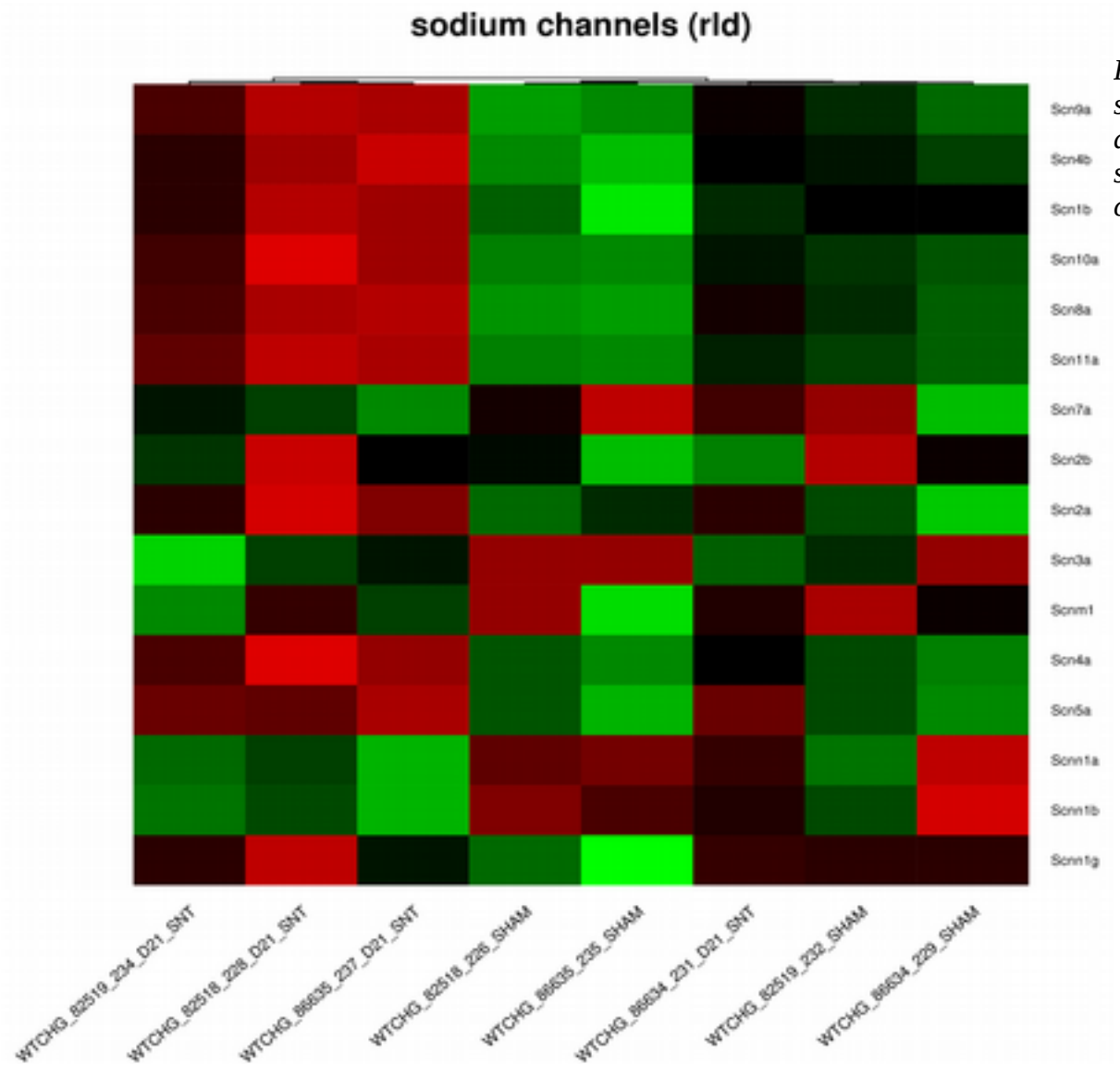


Figure 10: Expression patterns of voltage-gated sodium channels. Scn9a, Scn1b, Scn4b, Scn10a and Scn11a genes are consistently and significantly down-regulated in SNT samples comparing to sham.

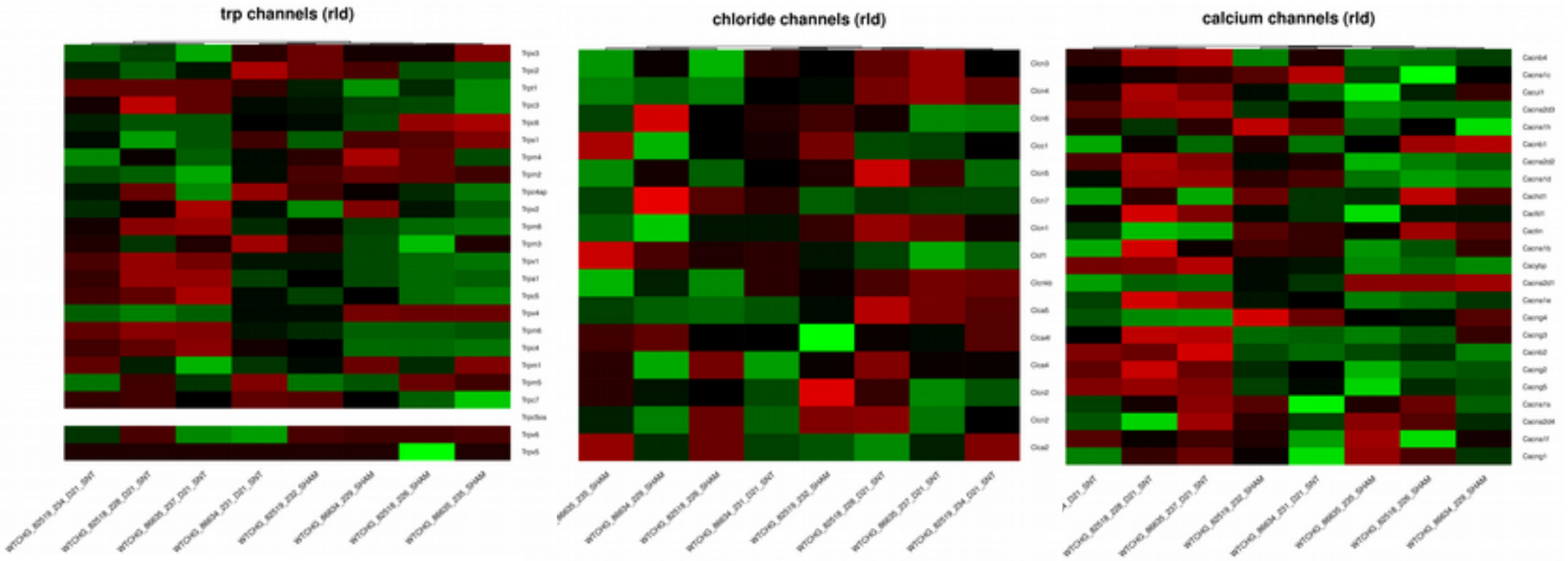


Figure 11: Heatmaps of ion channels. From left to right: TRP, Chloride and Calcium.

Identification of LncRNAs

After analysing the expression of known genes from the ENSEMBL annotation, we proceeded to identify novel LncRNAs and analyse their differential expression. We used the customised pipeline presented in the chapter Methods, section Reconstruct genes of putative LncRNAs, to generate a GTF file with predicted novel LncRNAs we calculated their coding potential and analysed their DE using DESeq2. We should note that due to the somewhat poor annotation of the rat genome we expected to identify a lot of un-annotated genomic loci predicted to be transcribed into putative LncRNAs as: 1. we do not have many annotated LncRNAs in rat 2. there could also be some not yet annotated protein coding genes in rat 3. due to the inferior quality of the rat genome assembly we expected some reads to be erroneously mapped in non-annotated regions of the genome, thus producing coverage which could then be identified as false positives.

As described in Methods, Identify expressed regions outside known gene models, in order to create an annotation at the gene level for novel LncRNAs we first discarded all reads mapping to known protein coding genes. Then we used the remaining to find Islands of Expression outside known gene models with a coverage of more than one. Subsequently, by selecting all *de novo* splicing junctions identified by STAR (Dobin et al., 2013) with more than 2 uniquely mapped or multi-mapped reads, we grouped together, trimmed those islands of expression and collapsed multiple overlapping transcripts to putative gene models of LncRNAs. As the rat genome is not that well annotated and we observed significant differences in protein coding gene models between RefSeq and ENSEMBL annotations, in order to avoid possibly un-annotated Un-Translated-Regions (UTR) flanking gene models or un-annotated exons, we extended gene models by 2000bp (Perkins, 2013). UTRs are transcribed but not translated, i.e. they do not have coding potential, and thus they can be indistinguishable from putative LncRNAs. The same can be true for small un-annotated exons. So we applied this screening method in order to avoid false positives due to the implications of some poorly annotated gene models.

After we imposed filtering, based on an expression level to be more than 0 reads across all of the samples in at least one condition, we got 7092 predicted LncRNAs in the rat DRG. We then filtered based on coding potential assessment carried out by the CPC support vector machine classifier (Kong et al., 2007). Kong et al found that CPC is more accurate in identifying coding transcripts and that CPC scores between -1 and 1 are considered to be in the grey zone. As we imposed several screening filters and we did not want to discard putative LncRNAs showing small coding potential we considered all transcripts in the grey zone to be putative LncRNAs. Another reason for considering this approach is that, as reported by (Ulitsky and Bartel, 2013), LncRNAs may be able to code for small peptides.

Regarding their exon structure, most of LncRNAs identified from our customised pipeline are bi-exonic and the distribution of their exon numbers heavily tails off after 4 exons, figure 12.

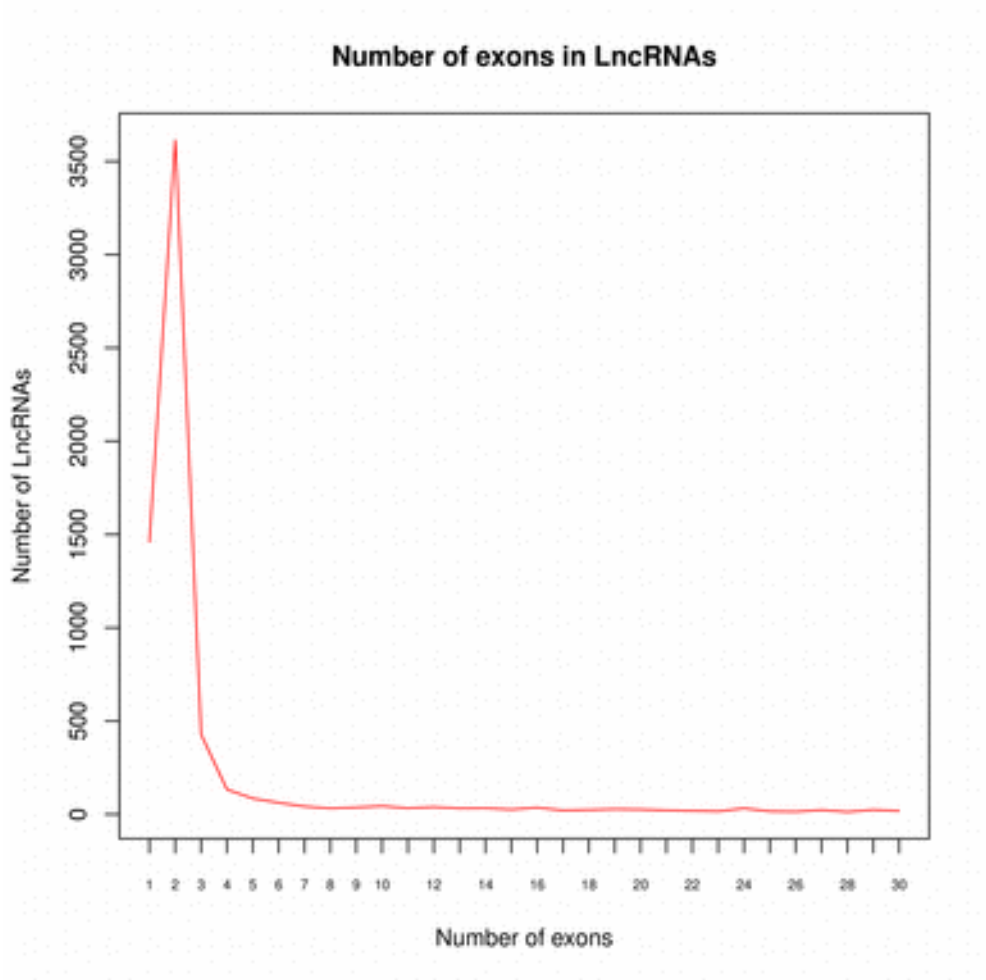


Figure 12: Distribution of exon numbers in the predicted LncRNAs

Expression of LncRNAs in rat's DRG

After assigning reads to the identified LncRNAs using HTSeq (Anders et al., 2015) and the *Intersection Not Empty* strategy we studied their expression consistency and strength relative to ENSEMBL genes. As seen in figure 13 LncRNAs were very lowly expressed, more than 5 times lower than ENSEMBL in their median value, figure 13.

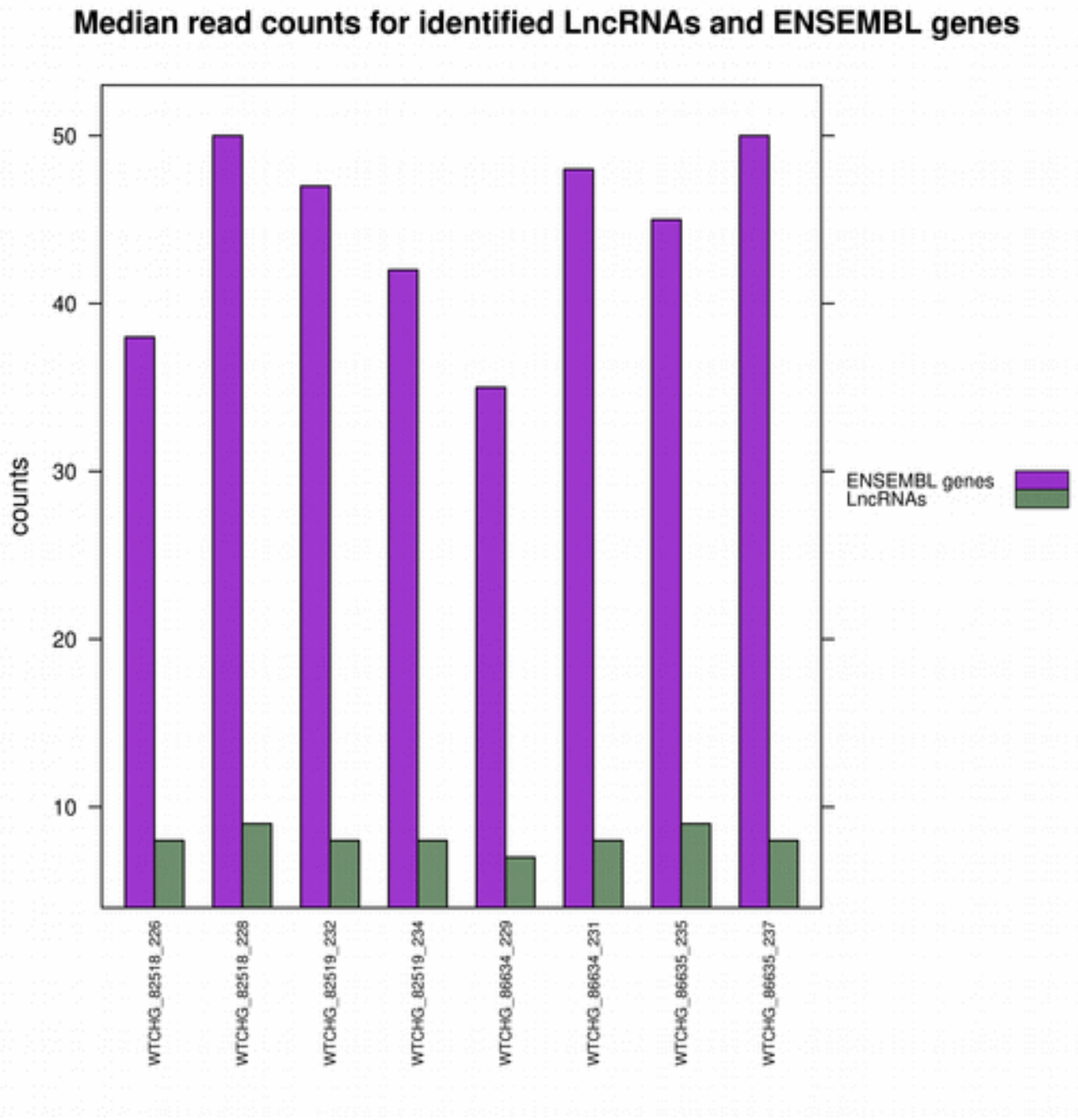


Figure 13: Median counts for ENSEMBL genes (violet) and predicted LncRNAs (green)

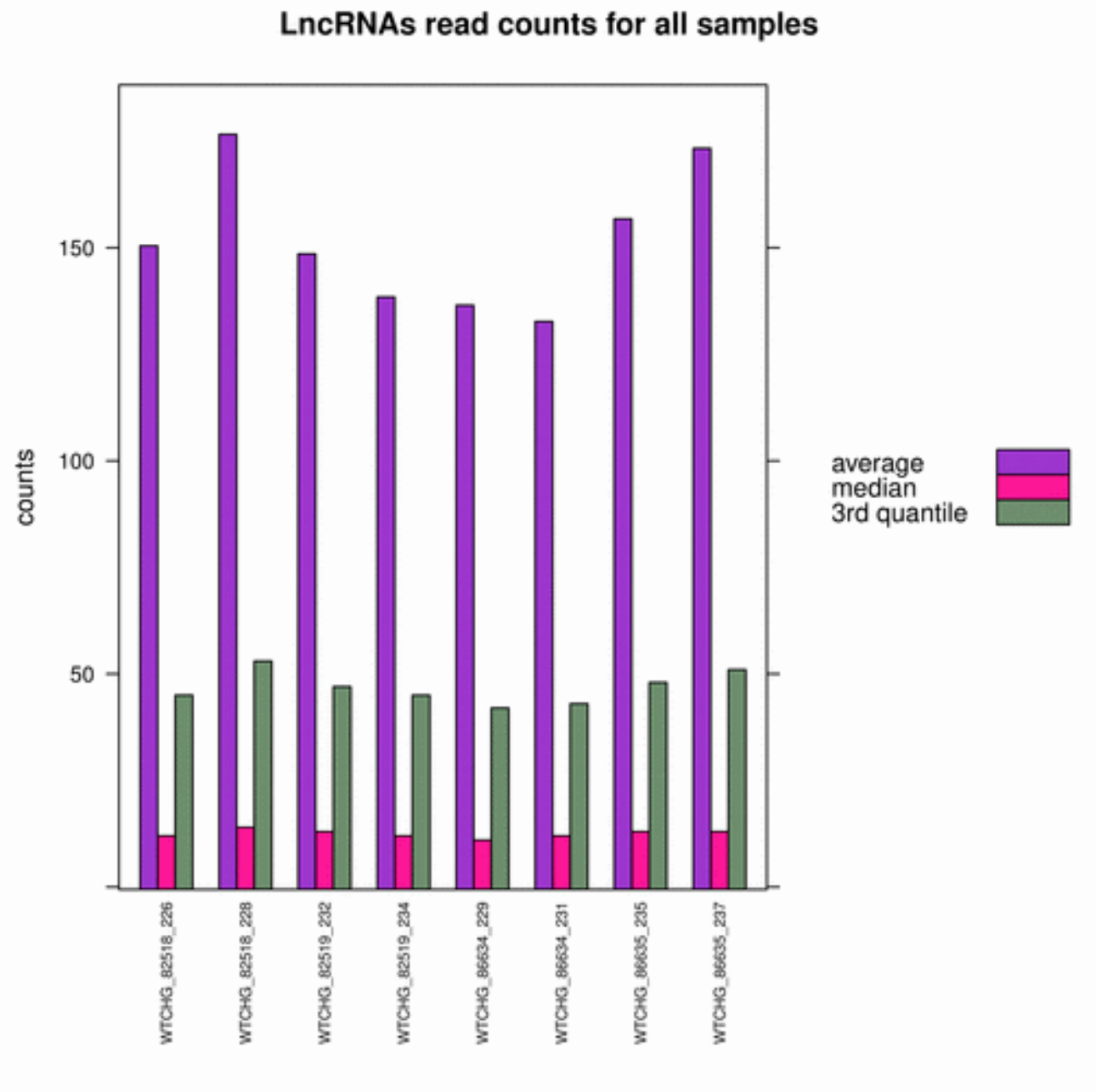


Figure 14: Distribution of read counts for predicted LncRNAs. Average (violet), Median (pink), 3rd Quantile (Green)

In order to further asses the quality of the LncRNA models identified we proceeded to assess the consistency of their expression across samples. As a metric for consistency we used Cook's distance, which stands for the difference in the coefficient of a linear model fitted for each gene, following a removal of a sample and refitting of the model. A large Cook's distance indicates inconsistent expression with spikes that could be due to

sequence artefacts, i.e. RNA-sequencing noise. We plotted the log10 of Cook's distance, thus negative values on the plot correspond to Cook's distances lower than 1, which is considered a very good value. As shown in figure 15, we observed a very consistent Cook's distance across samples.

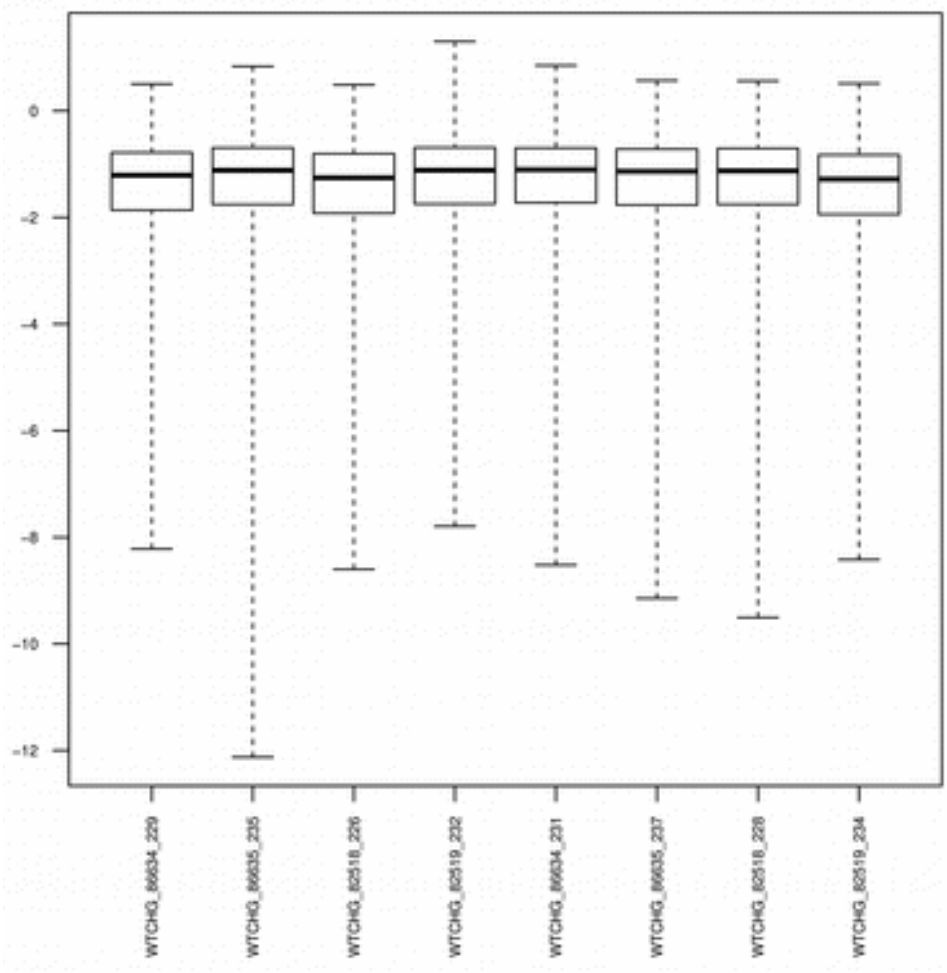


Figure 15: Boxplot of Log10 Cook's distances for all predicted LncRNAs for each sample.

Moreover, we identified the two known rat LncRNAs which are functionally important in pain and expressed in DRG. One is the LncRNA antisense of the Kcna2 gene (Zhao et al., 2013) and the other is the the LncRNA antisense of the Scn9a (Koenig et al., 2015). Both of them are antisense to known pain genes and Kcna2 is functionally associated with neuropathic pain. Our pipeline was able to identify both of these LncRNAs, table 3. In the paper identifying the Scn9a antisense LncRNA, the authors

report that it was not significantly dysregulated after pain models of peripheral neuropathy and the same was also true for the Scn9a protein coding gene. Moreover in the same publication both LncRNAs are found to slightly change their expression in the same direction, without reaching significance, in neuropathic pain and inflammatory models. Although in our case Scan9a and its antisense LncRNA were significantly downregulated.

Genomic Coordinates	Antisense pain gene ENSEMBL ID	Antisense pain gene symbol	Log2 fold change of LncRNA	Adjusted p.value of LncRNA	Log2 fold change of pain gene	Adjusted p.value of pain gene
chr2:229296258-229305952(-)	ENSRNOG00000018285	Kcna2	-0.73	0.17	-1.11	2.325E-06
chr3:59057403-59276093(+)	ENSRNOG00000006639	Scn9a	-2.36	1.116E-09	-1.086	1.445E-05

Table 3: LncRNAs antisense of Kcna2 and Scn9a gene.

Differential Expression of LncRNAs

We subsequently analysed the DE of LncRNAs using DESeq2. First we assessed whether our predictions carried any biologically relevant signal by inspecting how the expression of the identified LncRNAs separated our samples. A very good separation of sample according to condition, indicated that these putative LncRNAs are relevant to nerve injury, figure 16.

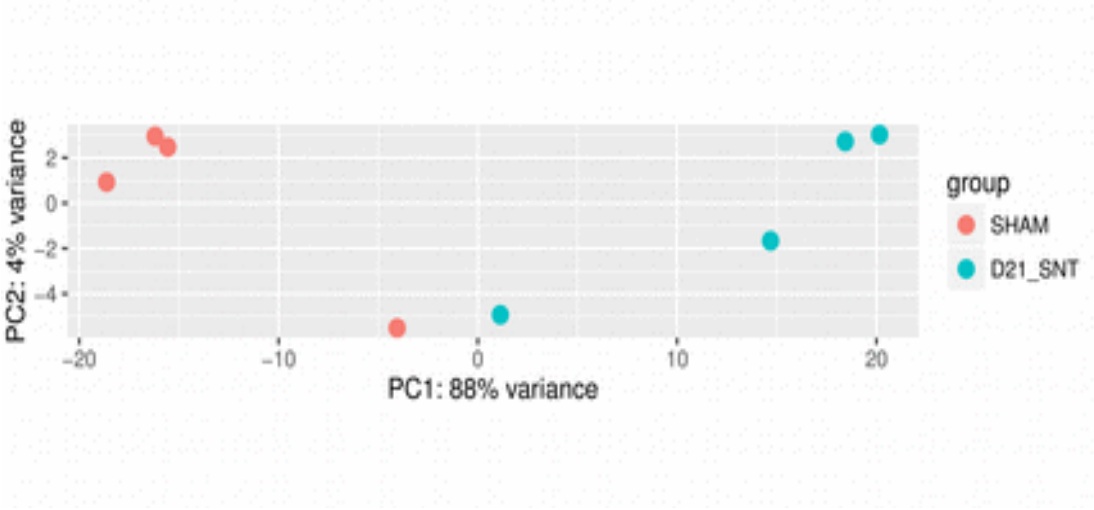


Figure 16: Samples' separation according to the expression of predicted LncRNAs.

After successful quality control we proceeded to the DE analysis. Using DESeq2 and the Wald test with the Benjamini–Hochberg correction to control false discovery rate and adjust p.values, we assessed the significance of differential expression of novel LncRNAs. As discussed above we had in total 7092 gene models of predicted LncRNAs with nonzero total read counts across conditions. Comparing rat SNT vs sham dorsal root ganglion we obtained the following results:

Predicted LncRNAs significantly differentially expressed (DE) with an adjusted p.value < 0.1 between SNT and Sham operated rats:

adjusted p.value < 0.05

LFC > 0 (up) : 274, 3.9%

LFC < 0 (down) : 577, 8.1%

outliers : 2, 0.028%

The very low number of outliers, suggested that there was no need for extensive moderation of log fold changes to accurately estimate differential expression. We should also note that the median counts across conditions were significant lower than that of ENSEMBL genes, a finding consistent with the literature for LncRNAs. For known ENSEMBL genes we obtained an average median across samples of 44.375 counts, for predicted LncRNAs we observed an average median of 9.5, almost a five fold change.

LncRNAs and pain-related protein coding genes

In order to infer the functional role of some of the predicted LncRNAs we examined them alongside their genomic context. A special class of LncRNAs are the antisense LncRNAs, lying on the opposite strand of protein coding genes. Antisense LncRNAs overlap some of the genomic regions of the gene model on the opposite strand, to varying extents and they can regulate the gene's transcription. As expected we identified two LncRNAs antisense of Kcna2 and Scn9a pain associated genes in rat's DRG. Then we investigated all antisense LncRNAs identified by the pipeline and studied them in the context of their expression pattern and function of the protein coding gene on the opposite strand.

In total, we identified 2300 antisense LncRNAs on the opposite strand of protein coding genes.

In rat DRG SNT vs sham we have 519 antisense LncRNAs significantly DE (adjusted p.value < 0.05). 77 are antisense of pain genes and out of these, 21 are significantly DE. 15 significantly DE antisense LncRNAs are antisense of significantly DE pain genes, table 4.

In this set there were important pain genes and ion channels. Moreover we identified a pair of protein coding gene and antisense LncRNA with opposite expression pattern in SNT vs sham. This expression pattern

could indicate pairs of protein coding genes / antisense LncRNAs where the antisense LncRNA acts as a competing endogenous RNA which silences the expression of the protein coding gene (Han and Jan, 2013). Kcnj9, a voltage-gated potassium channel, was reported to be associated with opioid and cannabinoid analgesia in mouse (Smith et al., 2008). In our study we identified an antisense LncRNA significantly DE with opposite expression profile to the Kcnj9 gene.

LncRNA	ENSEMBL ID	Genesymbol	lnc_lfc	lnc_pvalue	gene_lfc	gene_pvalue	cpc
chr1:177247980-177271718(-)	ENSRNOG00000017679	Cckbr	2.112658187	1.44658406879268E-06	1.9467315106	3.48587574213042E-07	-0.88056
chr10:40573511-40574610(+)	ENSRNOG00000012840	Sparc	0.693080439	0.0309146894	0.5219163833	0.0004593778	-0.972695
chr11:31735382-31864076(+)	ENSRNOG00000001575	Grik1	-1.2187665724	0.0008949842	-1.4425118869	0.0026368395	-0.940265
chr12:1148261-1148984(+)	ENSRNOG00000001090	Stard13	0.9088934954	0.0425706885	1.0436792285	0.0002948938	-1.18939
chr13:95223994-95228919(+)	ENSRNOG00000007645	Kcnj9	1.1049036718	6.56002103938819E-05	-1.2226296115	0.0003202419	0.459104
chr2:199111645-199115704(+)	ENSRNOG00000028589	Gria2	-1.5735337199	8.73053258903914E-05	-1.7163376183	8.44767579830593E-08	-0.97916
chr2:223469266-223471786(+)	ENSRNOG00000030019	Atp1a1	-1.5871758913	3.48442041275082E-05	-1.3611462919	0.000031922	-1.14917
chr3:59057403-59276093(+)	ENSRNOG00000006639	Scn9a	-2.357018731	1.42524585316199E-09	-1.0866688096	1.44524593696598E-05	-0.815863
chr3:127260488-127283308(-)	ENSRNOG00000014152	Kcnip3	-1.5362755276	0.00000623	-1.6266979042	4.14361222986572E-05	-1.15812
chr4:9636853-9646624(-)	ENSRNOG00000021441	Reln	1.2405189319	0.0029984059	0.7327351616	0.0018675684	-1.30911
chr4:162569914-162651062(+)	ENSRNOG00000005615	Gadd45a	0.9741379114	0.0011496768	1.0547848024	0.0022037369	-0.434463
chr5:3719233-3765605(-)	ENSRNOG00000007354	Trpa1	-1.0607031702	0.0005619469	-0.8457661675	0.0203168171	-0.814652
chr5:154402173-154493191(-)	ENSRNOG00000013231	Ptafr	0.677320791	0.0271052464	0.6466799012	0.0313546671	-1.52982
chr8:32323519-32324313(+)	ENSRNOG00000047179	Aplp2	-1.6363414407	0.0115691932	-0.7368788555	4.68583620320112E-05	-1.12105
chr8:88667082-88711088(+)	ENSRNOG00000013042	Htr1b	1.1366503299	5.71176513352015E-05	0.3209479749	0.0322417882	-1.06466

Table 4: Significantly DE LncRNAs antisense of significantly DE pain genes

All LncRNAs antisense of pain genes in rat's DRG can be seen in Appendix 2.

Regarding intergenic LincRNAs we found that the ones that were in genomic positions distant of known ENSEMBL genes tend not to be significantly DE. This might indicate that functionally important LincRNAs tend to be closer to protein coding genes, figure 17.

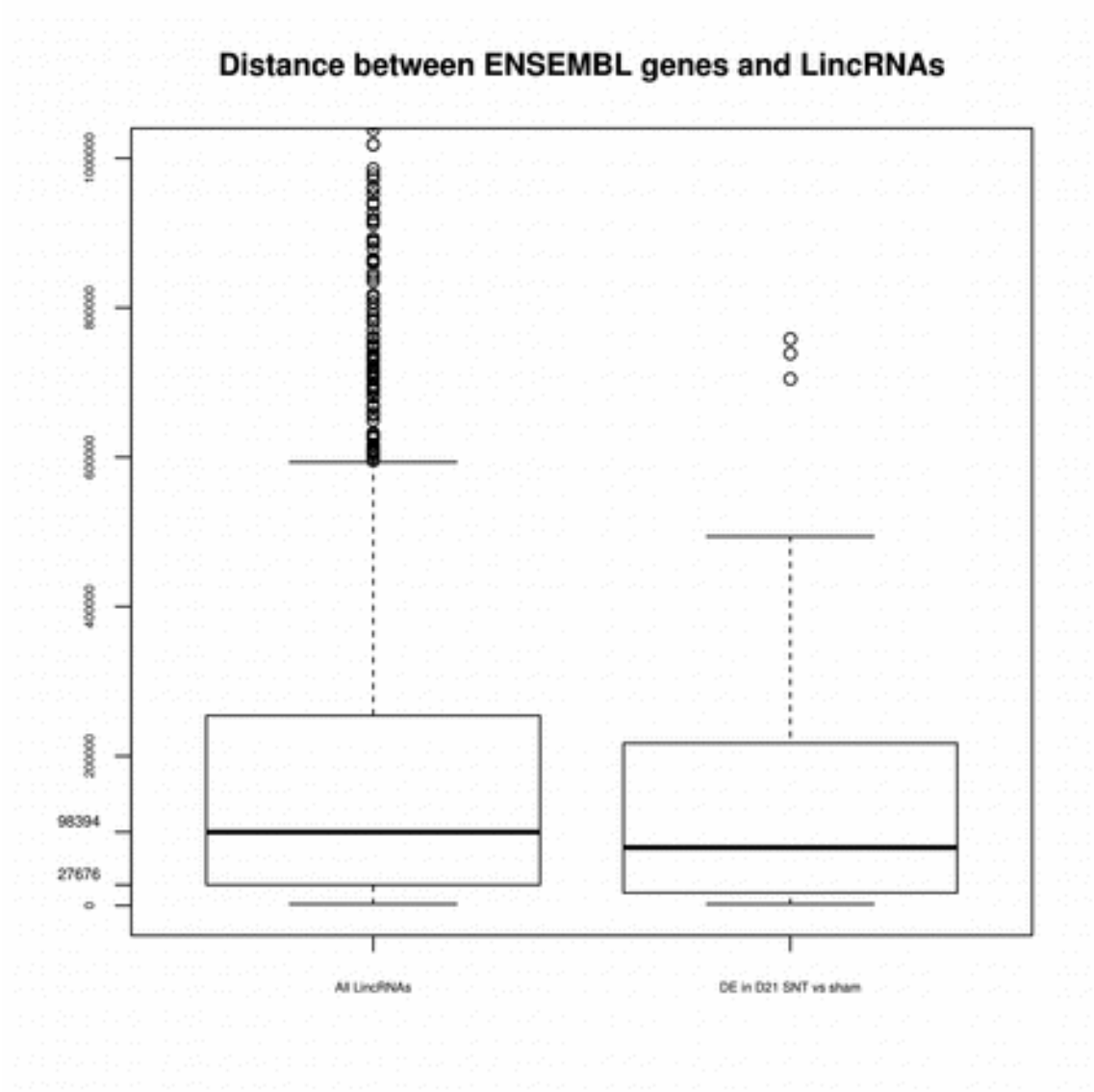


Figure 17: LincRNAs which are significantly DE (right-most boxplot) tend to be closer to ENSEMBL genes, with lower median (thick black line) and interquantile range (box height). Using all LincRNAs this difference in distance was significant a two-sample Kolmogorov-Smirnov test p-value =0.012.

51 LincRNAs were predicted to have a pain gene as their closest genomic neighbour. 11 of them were found to be significantly DE and 7

were DE with a significantly DE pain gene as their closest neighbour, table 6. Amongst them were Opioid receptors Oprm1 and Oprl1; Nefl which is functionally important for maintaining the neuronal caliper and intracellular transport; the transient receptor potential channel Trpa1; the serotonin receptor Htr1b; the neurotransmitter receptor Gria2 and the Disc1 which is related to neurogenesis. All these pain genes have a highly correlated DE LincRNA in very close genomic proximity.

As LincRNAs can be *in cis* regulators of gene expression either by inducing or silencing the expression of genes in close genomic proximity (Ulitsky and Bartel, 2013) we looked for pairs of protein coding genes and LincRNAs with highly correlated or anti-correlated expression patterns. All LincRNAs with a pain gene as their closest gene had positive correlations.

lincRNA	Distance	symbol	lnc_lfc	lnc_pvalue	gene_lfc	gene_pvalue	cpc	correlation	cor_pvalue
chr1:44859521-44866436(+)	-2315	Oprm1	-1.931967645	0.000117154	-1.7423528505	9.17347862170065E-06	-1.09013	0.9965496173	1.02427666792693E-07
chr3:180942195-180950495(+)	-2001	Oprl1	-0.8233498412	0.0312150005	-0.6804312538	0.0018394514	-1.2441	0.9499586396	0.0003016362
chr5:3736367-3743944(+)	20844	Trpa1	-2.0459213473	0.0007126908	-0.8457661675	0.0203168171	-0.960516	0.9623583176	0.0001296
chr8:88679840-88686412(-)	11861	Htr1b	0.9101055677	0.0263185643	0.3209479749	0.0322417882	-1.29518	0.7759879765	0.023593051
chr15:56397166-56423098(+)	-24859	Gfra2	-0.8000409826	0.0027784304	-0.3810290672	0.006427851	-1.03822	0.8690800403	0.0050735161
chr15:46543595-46565684(-)	-227842	Nefl	-1.202207748	2.41281227165041E-05	-1.4600252628	0.0007791237	-0.95852	0.9494949274	0.0003099898
chr19:68771054-68772954(+)	-2004	Disc1	0.7821331406	0.0152716877	0.5686654777	0.0147929284	-1.21019	0.873943157	0.0045462053

Table 6: Significantly DE LincRNAs with a significantly DE pain gene as their closest neighbour

As we observed generally high correlation between LincRNAs and their closest genomic neighbour we decided to assess significance of correlation utilising a randomisation / permutation approach using random pairing, see chapter Methods, section Calculate DE and associate expression profiles of putative LncRNAs and genes.

Given this approach, correlation coefficients larger than |0.75| were considered high. Thus all these pain genes had highly correlated expression pattern with the DE LincRNA in close proximity. We further discuss these results in the next chapter, in the context of the findings in mouse DRG after the Spared Nerve Injury (SNI) model.

Discussion

We have gained valuable insights into genes and the biological processes that are functionally important for neuropathic pain after peripheral neuropathy in rat. We confirmed that ion channels, genes associated with inflammation, neuron regeneration and development and opioid receptors are significantly dysregulated in rat's DRG under the SNT pain model. Moreover we established a strategy for identifying LncRNAs from RNA-seq and we calculated their DE alongside protein coding genes. We identified hundreds of predicted LncRNAs significantly DE between rats which underwent the SNT pain model and control samples.

We specifically looked at ion channels as GO biological processes related to ion channels were amongst the most significantly enriched processes in the set of DE genes. Moreover voltage gated ion channels are validated to be major contributors in pain and nociception. In the next chapter we found that two different mouse strains. One with high and one with low hypersensitivity after the pain model, had different transcriptional profiles of ion channels and pain genes. Regarding pain genes, are validated to be implicated in pain and helped us prioritize LncRNAs that may be functionally important to neuropathic pain. But more importantly, the fact that these genes had significantly different transcriptional profiles in rat was stressed also by the fact that these gene families (pain genes and in channels) had also different profiles between the high and low hypersensitivity mouse strains and that the high strain mouse is more similar to rat than the low pain strain. We present these results in the next chapter. Thus we confirmed that we had a very significant transcriptional response for all genes but also regarding pain genes and ion channels in rat and we also gathered evidence that this response is more similar to the response of mice having higher induced hypersensitivity after peripheral nerve injury.

In total, we identified 21 LncRNAs significantly DE antisense of pain genes. One of these pairs, the Kcnj9 gene and its antisense LncRNA had opposite expression pattern, which supports the hypothesis that antisense LncRNAs might silence the gene on the opposite strand.

Furthermore we observed, in general, high positive correlation between LincRNAs and their closest protein coding gene. 7 of these LincRNAs (table 6) were significantly DE and with highly and significantly positively correlated expression to that of pain genes, which were their closest genomic neighbour. These results fit with the hypothesis that either the transcription of the LincRNAs or the product of transcription, the LincRNA itself, induces the expression of pain genes *in cis*, or the transcription of the LincRNAs could be a by-product of the gene's transcription, also a moreover regulatory mechanism could be in function, the protein coding gene could regulate another gene using the LincRNA as an intermediate regulator. Thus we identified a stringent subset of LncRNAs which are consistently and sufficiently expressed in DRG, significantly DE in rats that underwent the pain model, and associated in terms of their genomic context and expression pattern to pain genes. This subset is comprised from LncRNAs for which we have evidence suggesting they are functionally important for the nerve injury response. This hypothesis requires further validation in the wet lab.

In the next chapter, we will further study transcriptional changes of genes and LncRNAs in DRGs of two mouse strains with high and low induced hypersensitivity after the SNI pain model of peripheral neuropathy and we will compare the results presented in this chapter.

Transcriptional changes of LncRNAs and protein coding genes in DRG of two mouse strains experiencing high and low induced hypersensitivity

Overview

To further investigate the underlying biological processes and mediators of neuropathic pain we exploited our customised pipeline presented, in Methods, in order to analyse Next Generation Sequencing data from mouse models of pain. Moreover we investigated whether mice of different strains, showing significantly different levels of hypersensitivity after pain surgery, show altered expression patterns of their protein coding genes and LncRNAs. Our hypothesis that certain protein coding genes and novel LncRNAs contribute to neuropathic pain are further assessed in the current chapter, using the advantages provided by the much better annotated mouse genome and a more comprehensive experimental design which assesses the impact of various factors including sex, strain and biological condition. We expect to find some of the conserved differentially expressed LncRNAs, identified in Rat Dorsal Root Ganglions (DRGs), to be also significantly DE between mice with induced neuropathic pain and healthy animals, as well as between strains with high and low hypersensitivity in the well induced neuropathic pain state. We used the Spared Nerve Injury (SNI) model of neuropathic pain. Surgeries and behavioural tests were carried out in Jeffrey Mogil's lab at McGill University, Montreal, Canada by Jean Sebastien Austin. Tissue extraction was carried out by Jean Sebastien Austin and me at McGill University. RNA extraction was carried out by John Dawes at David Bennet's lab, NDCN, NPP, Oxford University and library preparation and sequencing was carried out at Oxford Genomics, Oxford. All data analysis of behavioural and sequencing data was done by me.

Background

The Spared Nerve Injury pain model

In order to induce reproducible sensory dysfunctions like allodynia, hyperalgesia and spontaneous bursts of pain (Jaggi et al., 2011), in a period of time broad enough to allow extensive behavioural, clinical and molecular assays to be implemented, we used the Spared Nerve Injury model of peripheral neuropathy. As presented in the Introduction, in this model two of the branches of the sciatic nerve are axotomised and one is left untouched, spared, thus the model's name Spared Nerve Injury (SNI). In order for the model to be implemented correctly considerable caution is needed in order to not injure the untouched nerve. Different combinations exist, tibial and common peroneal axotomised and sural spared or the other way around. In our particular case we axotomised the common peroneal nerve and the sural nerve. Sparing only the tibial branch (figure 1) produces consistent and robust mechanical allodynia and hyperalgesia without increasing heat sensibility (Shields et al., 2003). All surgeries were carried out at Jeffrey Mogil's lab in McGill University by lab technician Jean Sebastien Austin. Surgical procedures followed published guidelines (Shields et al., 2003) according to which: After incision of skin and muscle, the common pereneal nerve and the sural nerve were tightly ligated with 9-0 silk suture. Then the two ligated branches were carefully transected in order to remove about 2mm of the stump of each distal nerve.

All mice received SNI surgeries on their left sciatic nerve branches. As the tibial nerve supplies sensory sensation for the sole of the foot, the paw and the skin of the paw, we assessed hyperalgesia after the SNI surgery on the left mouse paw. As a control, for each animal we also took repetitive measurements for the right paw, contralateral to the injury. To generate control, sham samples, the same surgical and anaesthetization procedures were followed, but instead of ligating or transecting the nerve branches these were just exposed.

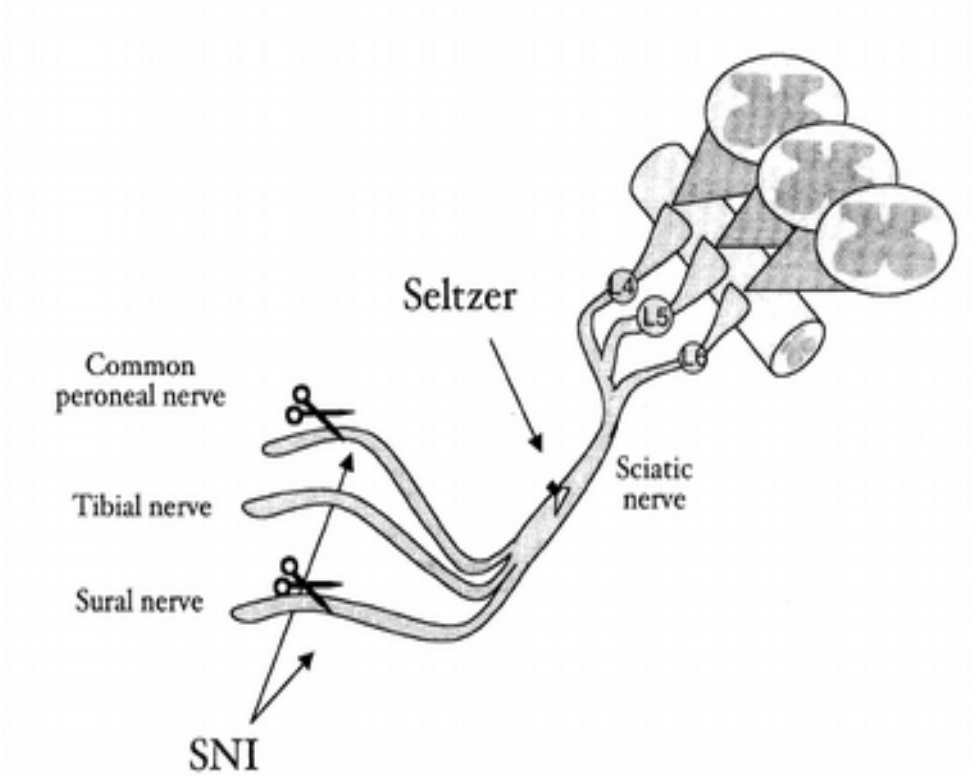


Figure 1: SNI pain model, the common peroneal ans sural nerve branches are axotomised. The tibial nerve is spared.

Behavioural tests

Behavioural tests were carried out using Von Frey filaments (see figure 2), one day, 7 days, 14 days, 21 days and 28 days after the surgery. These timepoints allow us to have a comprehensive view of the establishment and progress of painful neuropathy. In order to establish a base line four (4) different measurements were taken from the left (affected by the surgery) and the right (non-affected by the surgery) paw of the mice. Von Frey filament tests were repeated for the ipsilateral (left), figure 3, and contralateral (right), figure 4, paw. After the baseline had been established repeated measurements were taken with von frey filaments of different strength, in order to identify withdraw thresholds that increase mechanical sensitivity by 50%. The researcher who performed the tests was blind to the strain and condition of the animals. As seen in figure 3 and 4 we have significant and well induced hypersensitivity in the form of a consistent

decrease in the paw withdrawal thresholds for the ipsilateral paw of the SNI animals. Painful neuropathy is maintained from Day 1 to Day 28 and it increases with time. On the other hand we can see random natural variation for sham, i.e. control, samples both in ipsilateral and contralateral paw. In addition we can see no significant trend towards hypersensitivity in the contralateral paw.

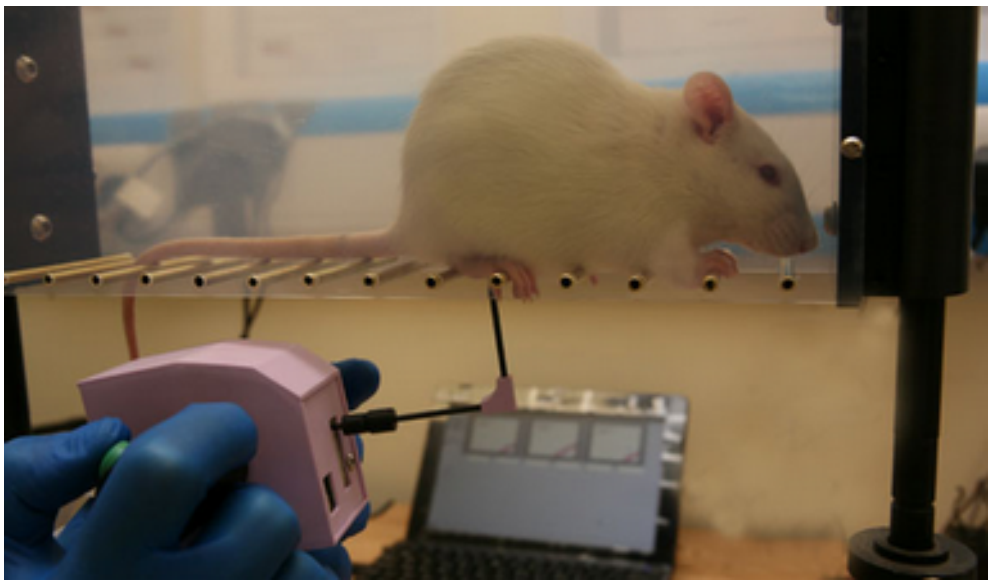


Figure 2: Mouse undergoing Von Frey filament testing for hypersensitivity in the pain lab

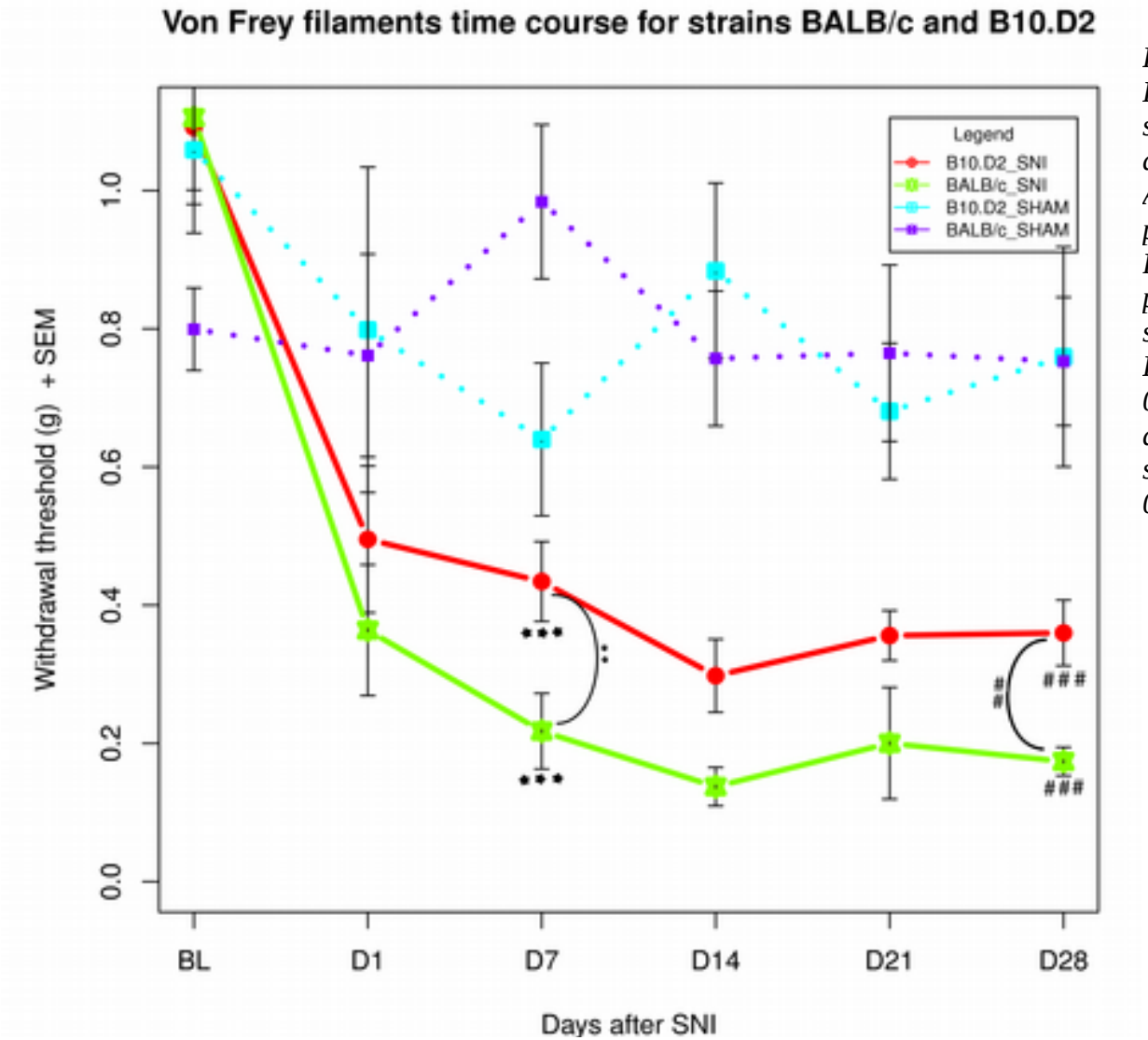


*Figure 3: Hindpaw withdrawal thresholds to Von Frey filament stimulation + SEM in grams. Both strains show significant induced hypersensitivity after the SNI surgery compared to baseline (two way ANOVA on Day 7 SNI p.value = 8.406e-06 *** and paired Welch t-test on Day 28 SNI vs baseline for BALB/c p.value = 7.855e-06 ### and for B10.D2 p.value = 9.581e-06 ###). BALB/c strain showed significantly different response to SNI surgery on Day 7 vs B10.D2 (two way ANOVA p.value = 0.008278 **) and induced hypersensitivity on Day 28 after SNI surgery for BALB/c vs B10.D2 strain was significantly different (Welch t.test p.value = 0.002395 ##)*

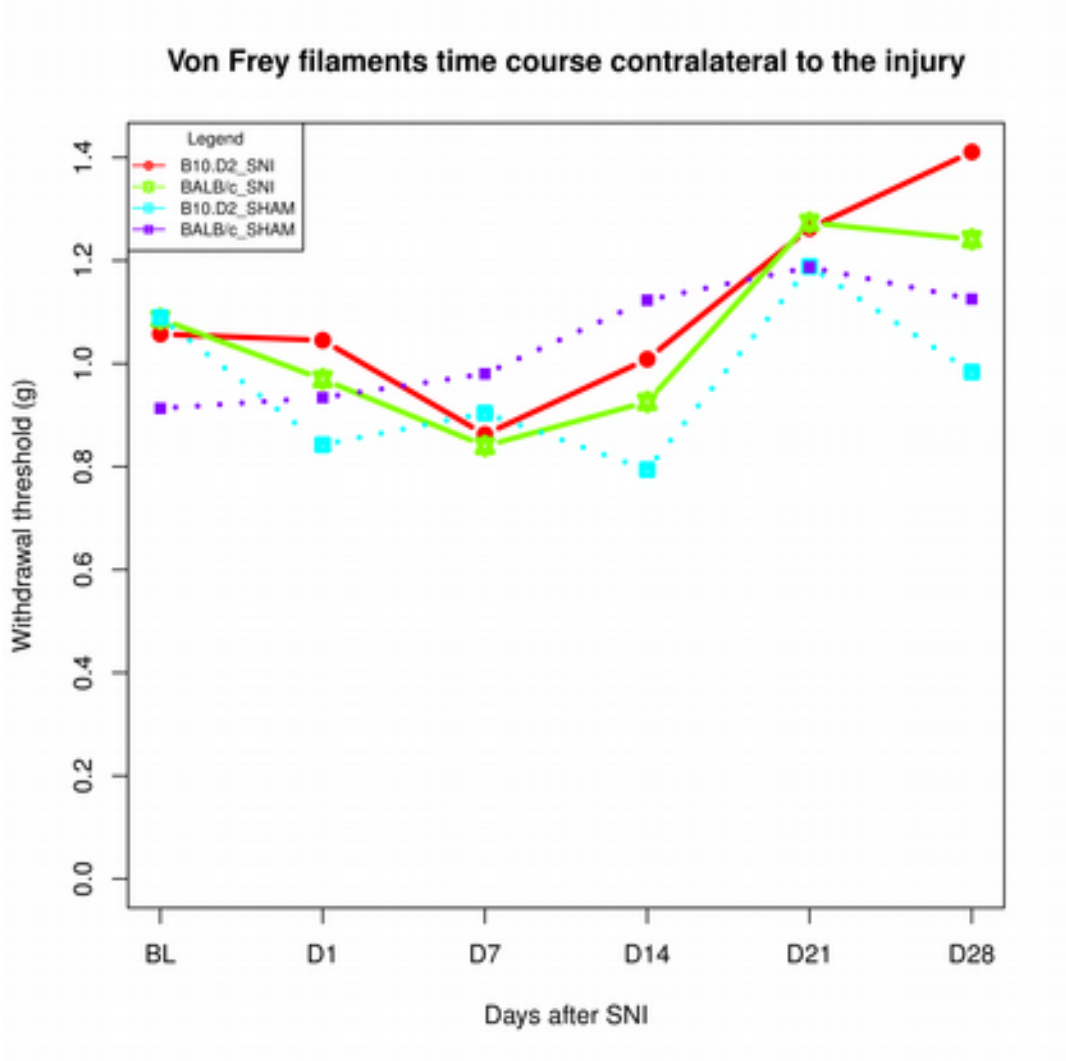


Figure 4: Withdrawal thresholds (in grams) for each mouse strain and timepoint. Right Ipsilateral paw. Solid lines are for the SNI mice, dashed lines are for mice that underwent Sham surgery only

Mouse strains and phenotypes

In the current study we used two different mouse strains, namely B10.D2 and BALB/c. B10.D2 is a recombinant congenic strain, and therefore carries a fraction of the genome of one strain together with the genetic background of another (Martin et al., 1992). In general recombinant congenic strains allow us to study the combined effects of minor genes, while on the other hand inbred strains allow us to effectively study the effects of major genes as most minor effects are masked. B10.D2 strain has been widely used in studies regarding the immune system response, thus it

is very relevant in studies that study neuropathic pain, pain with a significant inflammatory component. In terms of phenotyping, this strain can show prolonged allograft rejection, increased susceptibility to certain pathogens and impaired chemotactic responses of neutrophils

(“JAX® Mice & Services,” n.d.)The Jackson Laboratory, Bar Harbor, Maine. World Wide Web. URL: https://www.jax.org/, 6/2016). In pain studies B10.D2 mice have found to show low mechanical allodynia 5 and 7 days after SNI surgery (Sorge et al., 2012). In this study B10.D2 showed the lowest allodynia amongst 18 other mouse strains. Moreover this phenotype was associated with genetic variability of the P2RX7 receptor pore formation. B10.D2 mice, which carry the Leu451 allele, were experiencing far less allodynia and were insensitive to treatment that blocks the formation of P2RX7.

In contrast to B10.D2, the strain BALB/c (Eppig et al., 2015) is a commonly used inbred which presents susceptibility to the demyelinating disease upon infection with Theiler's murine encephalomyelitis virus. The most striking phenotypic characteristic of this particular strain is that it is Albino. Males are aggressive and will fight littermates. They are commonly used in cancer and immunology studies as well as in the production of monoclonal antibodies. They are known for being less susceptible to certain types of cancer (“JAX® Mice & Services,” The Jackson Laboratory, Bar Harbor, Maine. World Wide Web. URL: https://www.jax.org/, 6/2016). BALB.c mice show high allodynia after SNI surgery. In the study of (Sorge et al., 2012) BALB.c showed higher allodynia amongst 18 inbred mouse strains. On the other hand, it shows relatively low sensitivity, consistently, both in hot and cold heat stimuli (Mogil and Adhikari, 1999). Direct comparison between these two strains (Sorge et al., 2012) showed that the observed variability was due to the Leu451 allele of P2RX7, which is present in B10.D2 and blocks pore formation which in turn leads to low allodynia. On the other hand the Pro451 allele, which is present in BALB/c strain activates pore formation leading to more allodynia.

Our behavioural results, figure 3, are consistent with the literature and indicated a robust pain phenotype for both strains for a prolonged period of time after the SNI surgery. Moreover we found, as expected, that the *BALB/c strain showed significantly different response to SNI surgery on Day 7 vs B10.D2 (two way ANOVA p.value = 0.008278) and significantly higher induced hypersensitivity on Day 28 after SNI surgery vs the B10.D2 strain (Welch t.test p.value = 0.002395)*. Finally both strains have almost the same baseline mechanical hypersensitivity.

Dissections

After the SNI surgery and the behavioural tests, the mice were dissected. The animals were first euthanized by spinal cord dislocation. Then under the microscope muscle, connecting tissue and spinal cord were removed in order to expose the Dorsal Root Ganglions (DRGs). All dissections were performed on dry ice and RNase Decontamination Solution was used in order to prevent RNA degradation. For the purpose of our study we needed tissue from lumbar DRGs 4 and 5. In order to identify the L5 DRG we used the iliac crest or the tip of the iliad bone, “the first articular process more than 1mm rostral to the iliac crest” (figure 5) (Rigaud et al., 2008) was used as the landmark for identifying L5 in all samples. We harvested L4 and L5 DRGs from each animal and stored tissue in sterile eppendorf tubes using plenty of dry ice in cryobox containers, see figure 5.

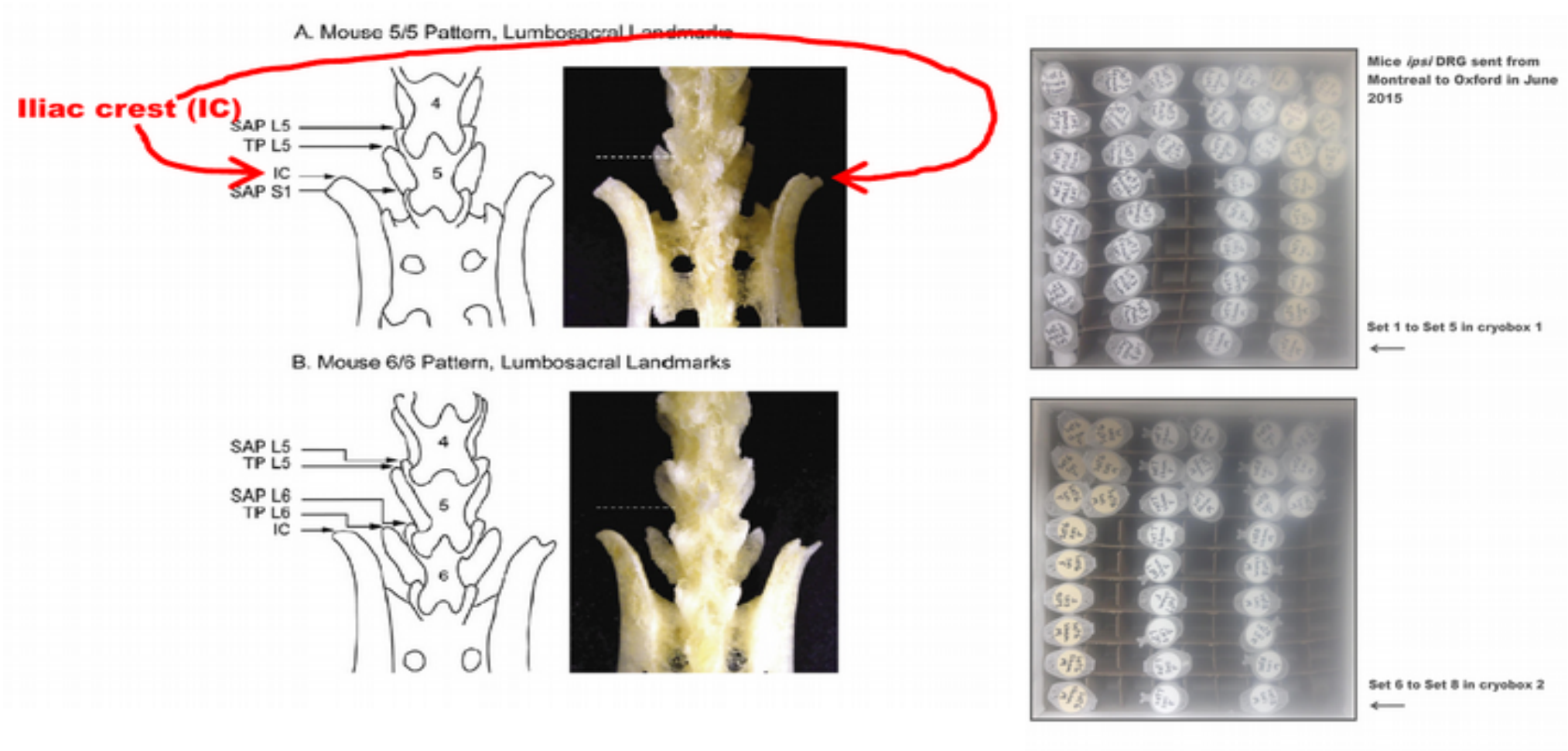


Figure 5: Left: Identification of the L5 DRG using the projection of the tip of the iliac bone.
Right: Eppendorf tubes with mice DRGs stored in cryoboxes.

RNA isolation and extraction

Dissected DRG tissue were sent from Montreal to Oxford in RNAse-free eppendorf tubes placed in dry ice at −78.5°C. RNA was then extracted by Dr. John Dawes according to the hybrid method of combined phenol extraction (TriPure, Roche) and column purification (High Pure RNA tissue Kit, Roche) (Bartus et al., 2016). According to this method two steps of RNA isolation and extraction take place subsequently, resulting in sufficient yield of high quality pure RNA. The concentration of RNA in the samples was measured using a nanodrop and proved sufficient for sequencing.

According to the hybrid method, DRG tissue was first homogenized in TriPure and then mixed with chloroform, following the phenol extraction method. After centrifuging the aqueous liquid, which stays on the top of the tube and contains the nucleic acids, was removed. This solution was then subjected to the column purification method. Following this protocol, the clear aqueous liquid was placed in Roche High Pure RNA tissue Kit columns and washed several times in order to purify the RNA. RNA was then extracted with the mRNeasy kit and all samples were subjected to on-column dnase digestion in order to prevent genomic contamination i.e. presence of DNA in the RNA samples.

Dataset

As seen in Figure 3, BALB/c and B10.D2 strains had the same baseline, significant and consistent increase in mechanical hypersensitivity after surgery and also presented the expected behaviour of low withdrawal threshold/high hypersensitivity, for BALB/c and high withdrawal threshold/low hypersensitivity, for B10.D2. Additionally we had consistent and expected variance for the sham groups. Based on these findings and on our initial hypothesis of identifying DE genes and LncRNAs between strains with high and low mechanical hypersensitivity (indicating high and low pain, after SNI surgery) we selected BALB/c and B10.D2 strains for RNA-sequencing. As described above we harvested 4 DRGs from each animal, L4 and L5. We had 24 animals, 12 of each strain. Two samples were

accidentally mixed together and destroyed during RNA-extraction, thus we used 22 samples for downstream analysis.

RNA-Seq and library preparation

The selected 22 samples were sent for sequencing to Oxford Genomics Center, High Throughput Genomics. Total RNA was provided to the sequencing center, and the ribodepleted fraction of it was selected for further sequencing. Ribodepletion excluded the ribosomal RNAs, see Introduction/RNA-sequencing. The ribodepleted RNA was then converted to cDNA using the strand-specific (deoxy-UTP strand-marking protoco) dUTP protocol which is the leading protocol for strand-specific synthesis of cDNA (see Introduction chapter, section RNA isolation and library construction).

Aligning RNA-seq reads to genome

Oxford Genomics produced FastQ sequencing files which encode quality metrics following the Sanger standard, i.e. Sanger qualities, using the standard Phred score (Ewing and Green, 1998) to asses the probability that the corresponding base call is wrong. Sequencing was done in five sequencing lanes producing five technical replicates per sample. In general all these lanes gave high yield, consistent GC content, consistent and expected sequence insert between the paired-end adapters and high quality base calling (see table 1). Duplicates were low, approximately 9% for all 5 technical replicates of the 22 samples.

Lane	% GC	% GC_{mapped}	σ_{pos}(%GC)	insert ± MAD	% exonic	% exon cov'ge	%N	max_{pos} %N	%lowQ	$\%lowQ_{end}$	avgQ
1.1	48.3 ± 9.6	47.8 ± 9.4	4.45	197 ± 59	20.8	59.9	0.0	0.3	0.0	0.0	34.0
1.2	48.3 ± 10.3	47.8 ± 10.0	2.61	195 ± 59	21.8	61.6	0.0	0.6	0.0	0.0	31.0

Lane	% GC	% GC_{mapped}	σ_{pos}(%GC)	insert ± MAD	% exonic	% exon cov'ge	%N	max_{pos} %N	%lowQ	$\%lowQ_{end}$	avgQ
2.1	48.1 ± 9.5	47.7 ± 9.4	4.46	191 ± 54	20.7	58.6	0.0	1.1	0.0	0.2	33.9
2.2	48.2 ± 10.2	47.6 ± 10.0	2.62	190 ± 53	21.7	60.3	0.0	1.0	0.0	0.1	30.8
3.1	48.2 ± 9.6	47.7 ± 9.4	4.45	192 ± 54	20.7	58.7	0.0	1.0	0.0	0.2	33.9
3.2	48.2 ± 10.3	47.6 ± 10.0	2.62	190 ± 54	21.7	60.4	0.0	1.1	0.0	0.1	30.9
4.1	48.2 ± 9.6	47.7 ± 9.4	4.45	192 ± 55	20.7	58.8	0.0	2.4	0.0	0.1	33.9
4.2	48.2 ± 10.3	47.7 ± 10.0	2.61	191 ± 55	21.6	60.5	0.0	1.3	0.0	0.1	30.7

Lane	% GC	% GC_{mapped}	σ_{pos}(%GC)	insert ± MAD	% exonic	% exon cov'ge	%N	max_{pos} %N	%lowQ	$\%lowQ_{end}$	avgQ
5.1	48.1 ± 9.6	47.7 ± 9.4	4.46	192 ± 55	20.7	58.4	0.0	0.4	0.0	0.0	32.5
5.2	48.2 ± 10.3	47.5 ± 10.0	2.58	191 ± 55	21.6	59.9	0.0	0.1	0.0	0.0	30.0

Table 1: Quality controls for all sequencing lanes. 1st lane, lanes 2,3,4, lane 5.

Next we mapped all of the sequencing samples using the spliced STAR-aligner (Dobin et al., 2013) to the latest mouse genome (mm10) downloaded from the ENSEMBL genome browser. Reads were aligned for each sequencing lane in parallel and then the sorted BAM files were merged with their respective technical replicates, from the same sample, in order to produce 22 merged, sorted and indexed BAM files. The mapping percentage was very good, with a mean across samples of 88.05% and all samples above 85% except for the Sample66_BALB.c_SNI_F with 73.1%. These numbers are considered very good and indicate samples of high quality.

We collapsed technical replicates before counting any reads, in order to achieve the best coverage possible in lowly expressed areas of putative LncRNAs. Then we counted reads for known genes using the ENSEMBL annotation. Again this was done in parallel, using HTSeq (Anders et al., 2015) with the "*intersection not empty*" strategy (see Introduction, section Analysing RNA-seq data). We also inverted reads' strand for counting as dUTP library produces reads with inverted strandedness. *Intersection not empty* counting strategy first calculates a disjoint version of the annotation and thus ensures that reads mapped to overlapping features are counted only once. In order to assign reads to known genes we used ENSEMBL's (Flicek et al., 2011) genome-build GRCm38.p4, based on genome version GRCm38, last updated in December 2015. We assigned counts on the gene level grouping together multiple transcripts derived from the same gene.

Experimental Design

As described above we had two strains for downstream analysis, one with high pain (BALB/c) and one with low pain (B10.D2), 22 samples in total. Samples are balanced regarding the strain, 11 samples from the

B10.D2 strain and 11 from the BALB/c strain. 12 samples underwent sham surgery and 10 the SNI pain model. Ideally samples should have been stratified for every experimental factor, but as we have enough replicates for each distinct combination of strain and condition (n > 3) we can efficiently estimate dispersion within conditions of interest and assess the significance of differences between conditions of interest.

In this experiment we had two conditions of interest, SNI vs Sham surgery and two types of strain, B10.D2 vs BALB/c strain. In the context of the generalized linear model formulation, we had 3 coefficients of interest, condition (with levels sham or SNI), strain (with levels BALB/c or B10.D2) and sex (with levels Male or Female). First we assumed that sex coefficient can be added uniformly across samples, as the differences between sexes are not of primary interest. Nor we are interested in a strain or condition effect that affects only one sex. We would like to adjust for any differences occurring due to the sex difference in order to effectively assess genes dysregulated due to changes in strain and condition. Thus we used an additive or blocking design for gender. Regarding the coefficients of primary interest, strain and condition, we assumed that there is some interaction between them. Thus the coefficient for condition is not the same across strains and vice versa. In this way we investigated all possible levels of the factor condition (sham or SNI) for each strain separately. Moreover we tested if and how different is the coefficient of condition between strains. For a more detailed discussion regarding comparing conditions see Methods, section Comparing conditions using Generalized Linear Models (GLMs). We used a nested interaction model for strain and condition which allowed us to test all possible interactions between them. The GLM we fitted for each gene is of the following form:

~ *sex + strain*condition*.

Then we tested if the coefficient of condition is significantly different from zero for:

1. The baseline strain BALB/c (the main effect, high pain)

2. The low pain strain B10.D2 (the main effect plus the interaction term representing only the difference of the coefficient for B10.D2 strain)

3. The interaction term, only assessing whether the log fold changes SNI vs sham are significantly different between B10.D2 vs BALB/c.

Further quality control

Before the analysis we assessed how reads were distributed on distinct features (genes, LncRNAs) of the annotations (gene sets in the form of GTF files) used and how consistent was the expression pattern of genes across samples. We had used two types of annotation, one is the latest ENSEMBL/GENCODE annotation the other is a customised annotation of predicted LncRNAs identified from our customised pipeline.

Then we examined how RNA-seq reads were distributed on the annotations' features. RNA- seq reads carry a distinct header coded in the hexadecimal system. This header flags reads according to their mapping and quality attributes. Reads can be uniquely mapped, mapped in more than one position on the genome but with only one optimal mapping position and mapped in multiple positions with equal probabilities (multi-mappers). As a consequence these multi-mapped reads are aligned in several genomic regions and produce multiple non-unique alignments, we assessed how many they were and how they could affect gene expression analysis by measuring the number of non-unique alignments they produced. Moreover, as transcript abundance estimation involves counting of the reads overlapping distinct features of the annotation, reads can overlap one feature alone or they can overlap more than one feature (ambiguous reads). Ambiguous reads can only be examined in the context of a specific annotation as they emerge as a combination of read mapping and the annotation structure. Ambiguous reads are reads that can be aligned to the genome, but the position that they are mapped happens to overlap more than one genomic feature. As we used two different types of annotations, i.e. a standard ENSEMBL and RefSeq annotation and a customised annotation for predicted novel LncRNAs, we examined ambiguous reads for both

annotations. A particular sample with more ambiguous reads and multi-mappers than the other samples may indicate genomic contamination, RNA degradation or poor sequencing quality.

Figure 6: Number of non-unique alignments due to multi-mappers (violet) and ambiguous reads (pink) for each sample. The left plot presents the distribution for ENSEMBL genes; the right plot for predicted novel LncRNAs. Sample 59 has a significant spike of multi-mapping and ambiguous reads. The same is also true for Sample 66. Our customised annotation of predicted LncRNAs gave very good results in terms of ambiguously assigned reads, having near zero for all samples, in the counting step.

As seen in figure 6, Sample 59 has significantly more non-unique alignments due to multi-mapped reads than the other samples and produces much more ambiguous reads for both annotations. As the amount of multi-mappers, i.e. ambiguously mapped reads on the genome, is solely an attribute of the sample and does not depend on the annotation, it suggests that Sample 59 suffers from poor quality. Moreover, most probably due to the high number of multi-mappers, it produces more ambiguous reads, i.e. reads overlapping more than one feature of a genomic annotation, regardless the of the annotation. Sample 66 had also a significantly higher number of multi-mappers and also produced higher ambiguous reads in the counting step. We should note here that our customised annotation of predicted LncRNAs had a consistently low number of ambiguous reads, similar to the ENSEMBL genes, indicating good quality of annotation with no randomly overlapping features that could produce a lot of ambiguous reads.

Next we examined the distribution of the gene's Cook's distances in all samples. Cook's distance is the difference in the coefficient of a linear model if we remove a sample and refit the model. Thus it assesses the consistency of expression of genes across samples. Higher Cook's distances observed in a particular sample, indicate that genes in this sample are expressed in an outlying fashion. High Cook's distances of particular genes indicate that these genes are not consistently expressed but they have rather serious spikes in their expression across samples, thus this is a way to detect outliers. The distribution of Cook's distances across all samples can be seen in figure 7. Again Sample 59 showed higher Cook's distance than all other samples, followed by Sample 66. Sample 66 also had the highest interquartile range, i.e. spread between the 1st and the 3rd quartile, indicating higher dispersion of gene expression than all the other samples.

Based on these findings that consistently indicated poor quality of these samples, plus on the fact that Sample 59 showed a significantly lower percentage of mapped reads we decided to exclude both Sample 66 (BALB/c SNI) and Sample 59 (B10.D2 Sham), as these samples had significantly inferior and outlying quality metrics. Given that these samples

are from both strains and conditions we did not change our experimental design and we ended up with 10 BALB/c and 10 B10.D2 mice.

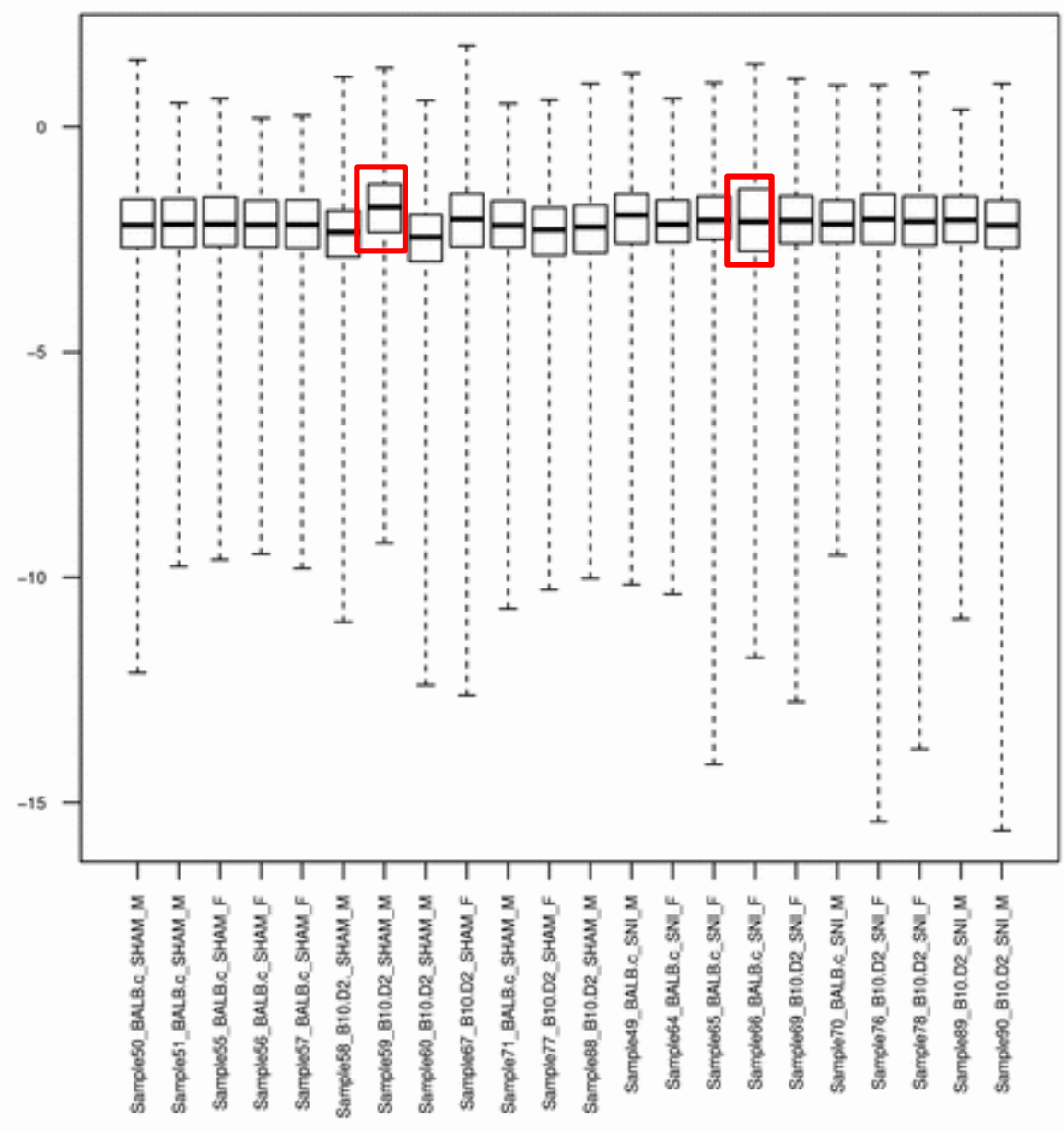


Figure 7: Boxplots of ENSEMBL genes' Log10 Cook's distance across all samples. The thick black line represents the median. Sample 59 has higher Cook's distance than all the other samples. Sample 66 has the higher interquantile range represented by the height of the boxplot. Gene expression is very consistent across all other samples

Results

Differential Expression analysis of known genes

Before analysing gene expression we clustered samples in an unsupervised way, blind to the samples' annotation, using hierarchical clustering according to the euclidean distance. In this way we assessed how samples of different conditions and strains were separated.

We first normalised and log transformed counts in a way that does not overestimate Log Fold Changes (LFC) for genes with low counts using the regularized logarithm transformation (rld) (Love et al., 2014). Then we proceeded to perform clustering and generated visualisations. Clustering of all 20 samples selected for downstream analysis indicated that transcriptional changes are more significant between strains, and that within strains there is a clear separation between conditions, figure 8. There is also separation between sexes in some extent.

Moreover if we separately inspect clustering of samples for each strain we can see that most of the samples clustered well within their respective family, except for one outlying sample for the BALB/c strain, figure 9. Principal components analysis of gene expression patterns for both strains showed that we have better separation between Sham and SNI samples for the BALB/c strain than for the B10.D2 strain, figure 10.

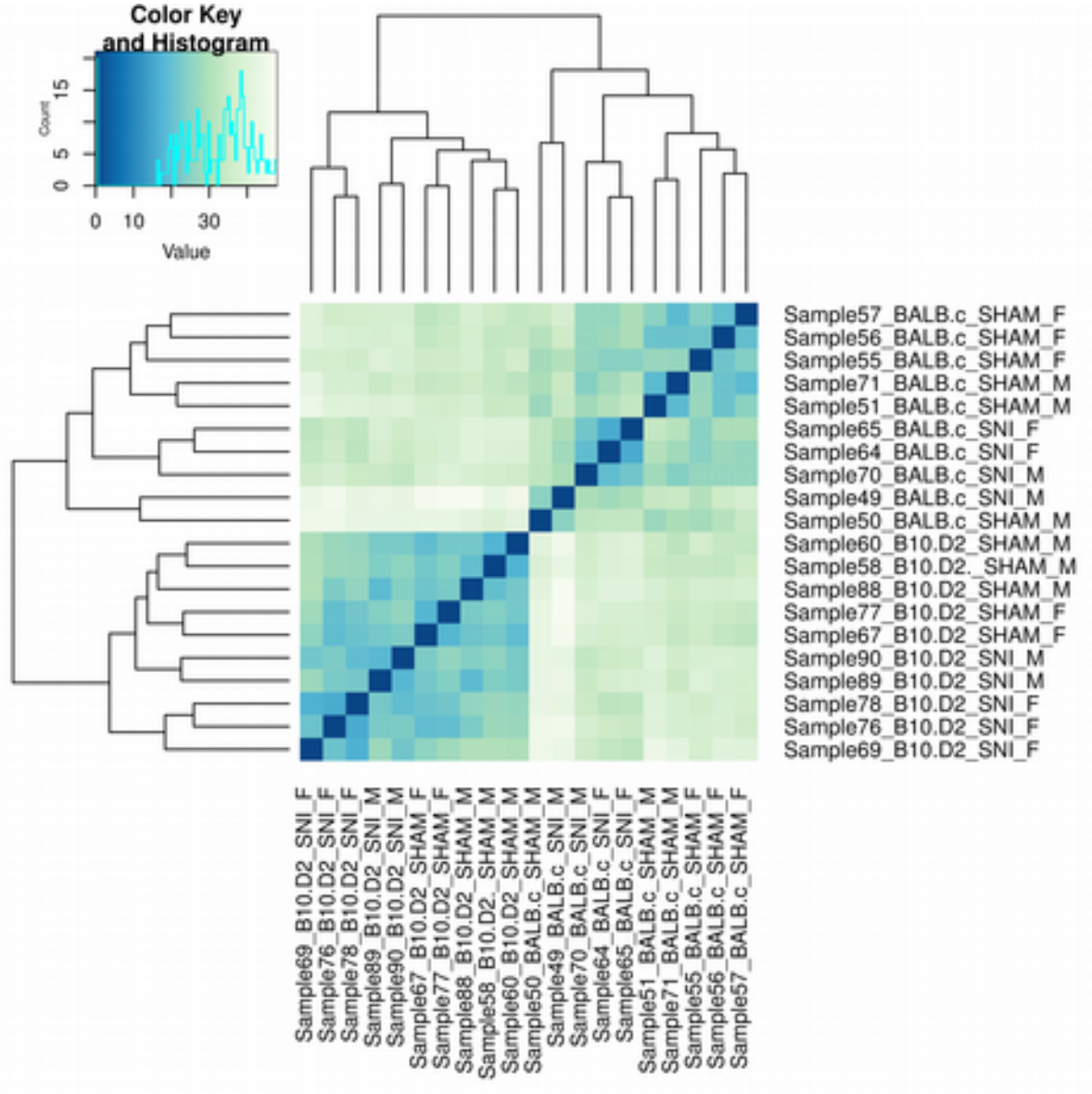


Figure 8: Hierarchical clustering of samples according to regularized log2 counts of ENSEMBL genes. Samples are clustered first according to strain and then according to condition. Genders are also separated within strain and condition.

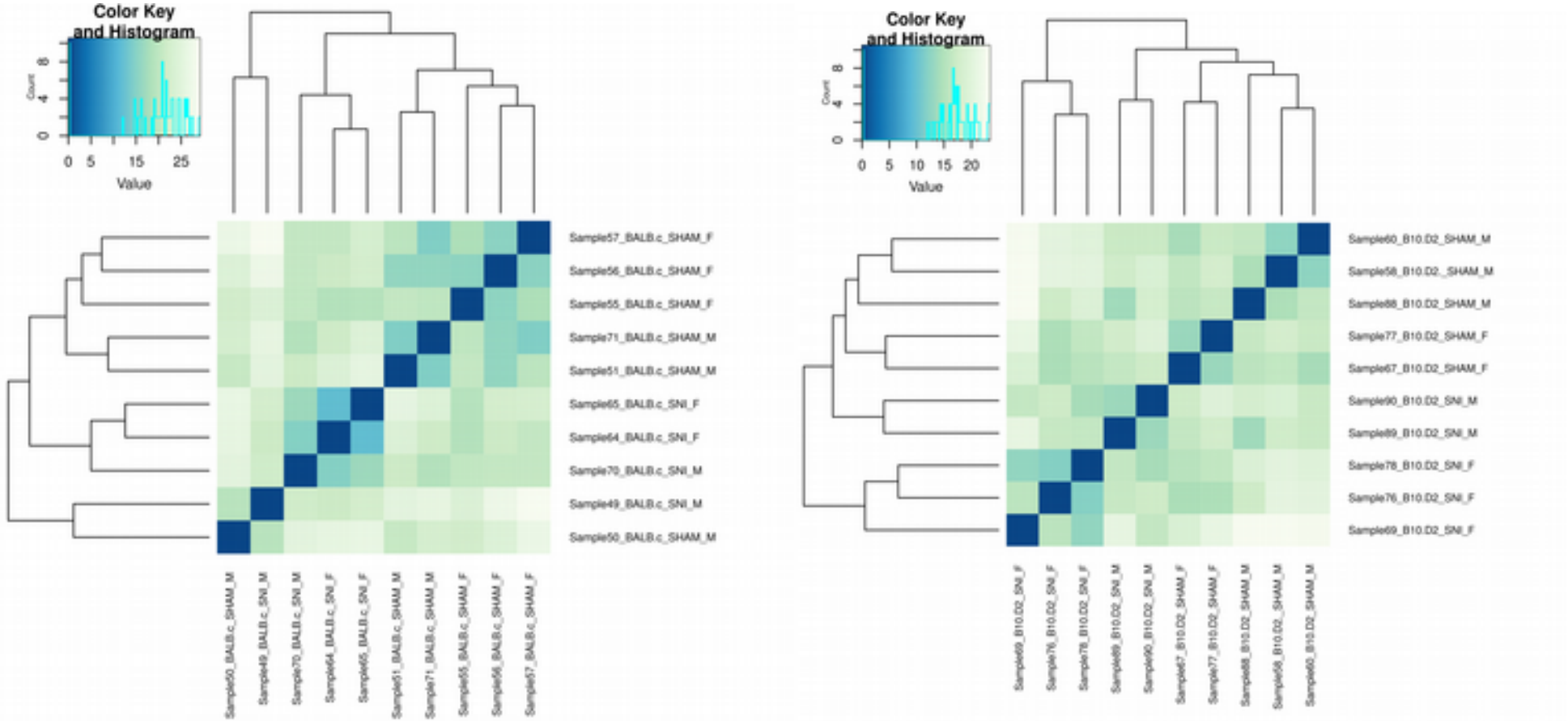


Figure 9: Clustering of BALB/c (left) and B10.D2 (right) sample according to regularized log2 counts of ENSEMBL genes.

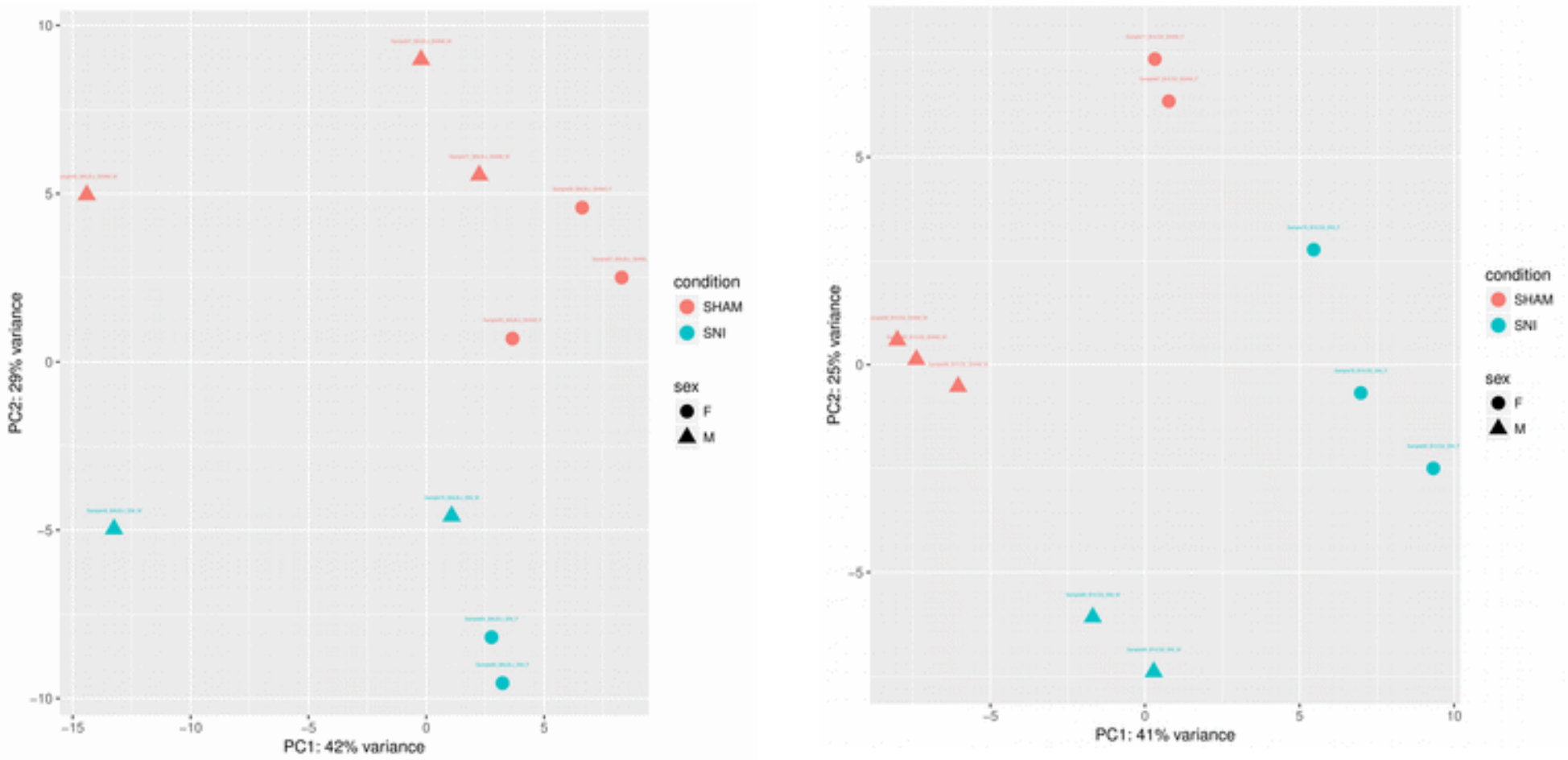


Figure 10: PCA for BALB/c strain (left) and B10.D2 strain (right).The top 50 ENSEMBL genes contributing to PC1 and PC2 are in Appendix 3.

Expression patterns of ion channels and pain genes

Being aware of the significance of ion channels in neuropathic pain (Basbaum et al., 2009) and having established very prominent changes of the expression level of ion channels and pain genes in rat DRG we used the same log transformed counts to analyse the transcriptional patterns of ion channels and pain genes across different strains and conditions.

Nociceptors express voltage gated channels in order to transmit pain from the peripheral nervous system. Moreover a set of known pain genes validated in knock-out studies from rodent models of pain in combination with a list curated from meta-analysis studies of pain (LaCroix-Fralish et al., 2011; Lacroix-Fralish et al., 2007), is enriched with genes coding for these channels as well as genes involved in inflammatory process, apoptosis and neurogenesis. Therefore we decided to specifically look at ion channels and known pain genes. We selected all potassium, calcium, sodium, chloride, transient receptor potential channels and pain genes and plotted their expression patterns in heatmaps and examined how these specific gene sets could separate samples from different strains according to condition.

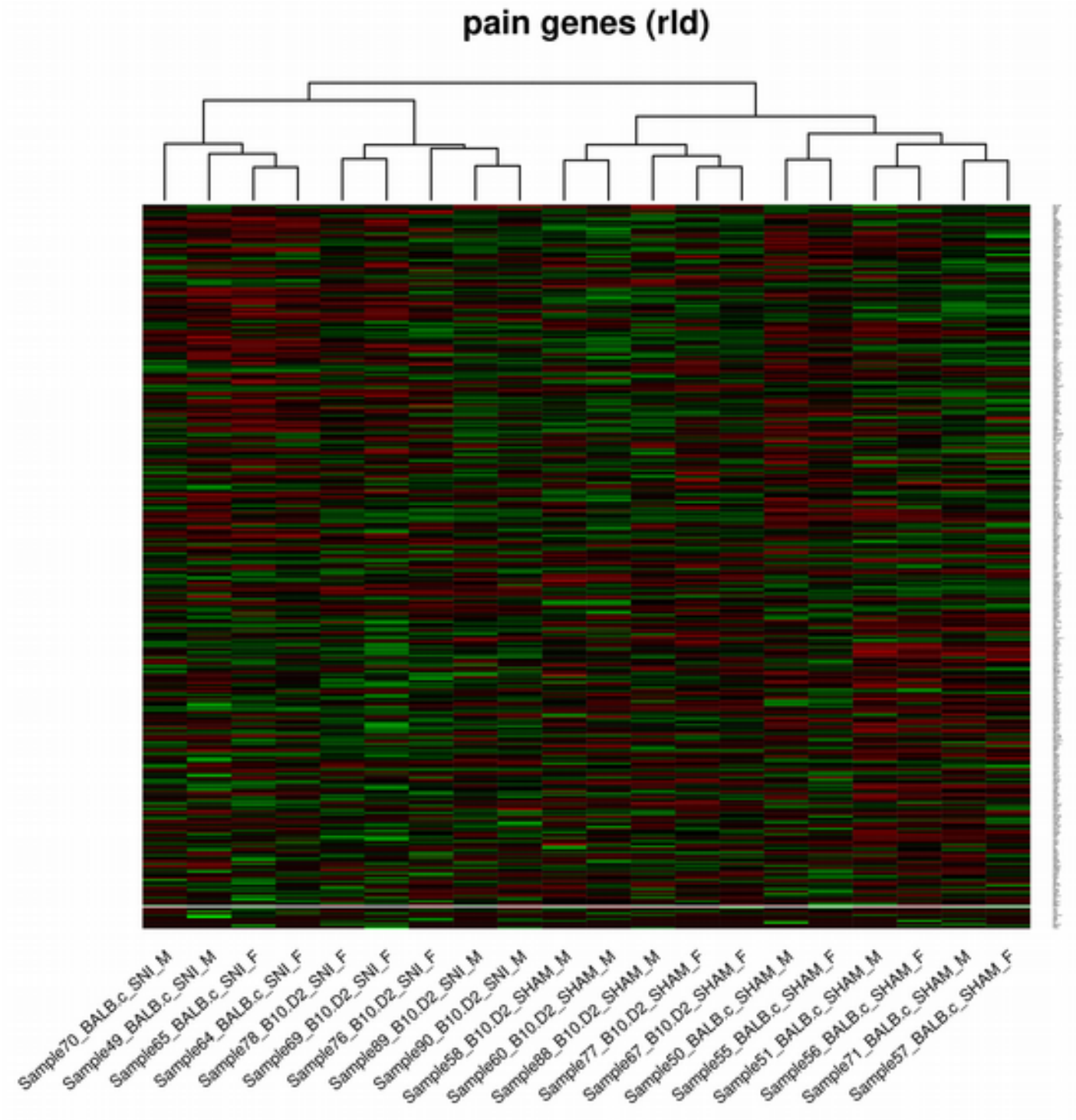


Figure 11: Expression pattern of pain genes based on rld transformed counts. Using only this subset of pain genes samples are perfectly clustered according to conditions and we can also see balanced down-regulation (top left) and up-regulation (bottom left) after the SNI surgery.

First, by examining the expression patterns of pain genes downloaded from the the pain genes database (Lacroix-Fralish et al., 2007) we found that all SNI samples were grouped together and the same was true

for all Sham samples, see figure 11. In addition the effect of strain was less prominent than in the whole ENSEMBL gene set. Samples were optimally separated according to condition (SNI vs Sham) based on the expression of pain genes, while based on the expression of the whole ENSEMBL gene set samples were optimally separated first according to strain and then according to condition. Moreover there was a distinct subset of genes upregulated after the SNI surgery and a distinct subset downregulated after the surgery. As it is expected when we selected only pain genes which were significantly (adjusted p.value < 0.05) DE in both strains, we observed a very clear separation between conditions as well as distinct groups of co-up-regulated and co-down-regulated genes, figure 12.

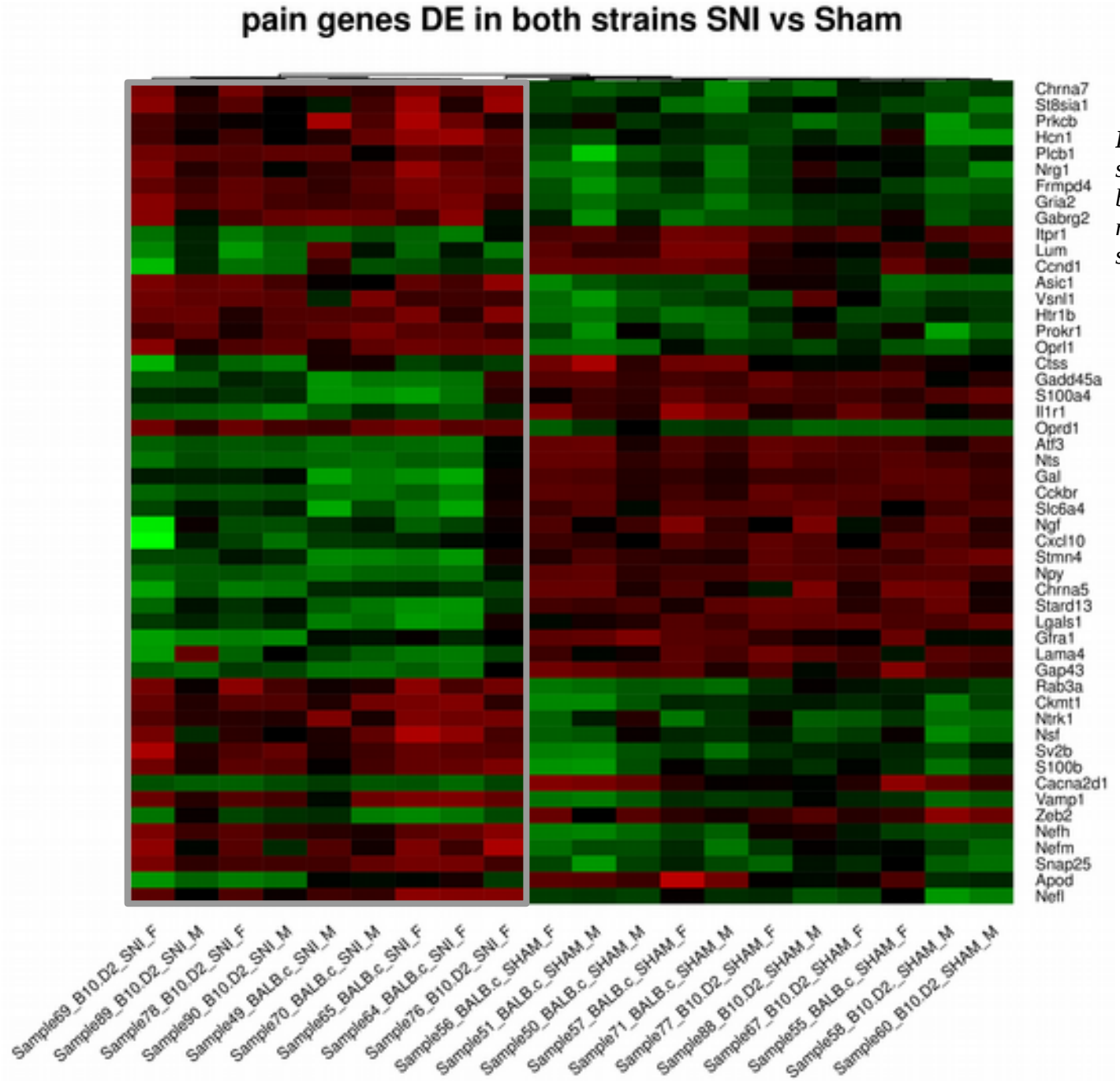


Figure 12: Common significantly DE pain genes in both strains. The grey rectangle denotes the SNI samples

To examine the differences between strains, we produced heatmaps for all the expressed ion channels in our dataset, for each strain separately. Interestingly we can see that expression of potassium and sodium voltage-gated channels provides optimal separation between animals with painful neuropathy and animals that have undergone sham surgery for the BALB/c strain. In BALB/c strain there were two distinct profiles of transcriptional changes regarding these channels, very similar to what we observed in rat. However genes encoding for these channels did not optimally classify samples from the B10.D2 strain, figure 13. Regarding B10.D2 strain, there were mostly three profiles which did not separate samples according to condition, except for chloride channels. Chloride channels could optimally separate samples according to condition for both strains; Thus the expression profile of ion channels for the high pain (BALB/c) strain is very similar to rat (figures 9,10,11 pages 103:105) while on the other hand the transcriptional response doe thw low pain strain (B10.D2) is different. TRP and calcium channels' expression pattern did not optimally separate samples according to condition for either strain.

The above observations suggest that BALB/c sample, which exhibits significantly higher induced mechanical hypersensitivity after the SNI surgery than B10.D2, had a more prominent pain signature at the molecular level mainly associated with the expression pattern of voltage-gated potassium and sodium channels. Moreover pain genes and especially those which were significantly DE in both strains (figure 12), comprised a set of genes with a robust dysregulation pattern that can separate SNI from Sham samples. The same was also true for chloride channels (figure 13). On the other hand potassium and sodium voltage-gated channels are more robustly dysregulated in the high pain strain BALB/c (figure 13) than in B10.D2 strain.

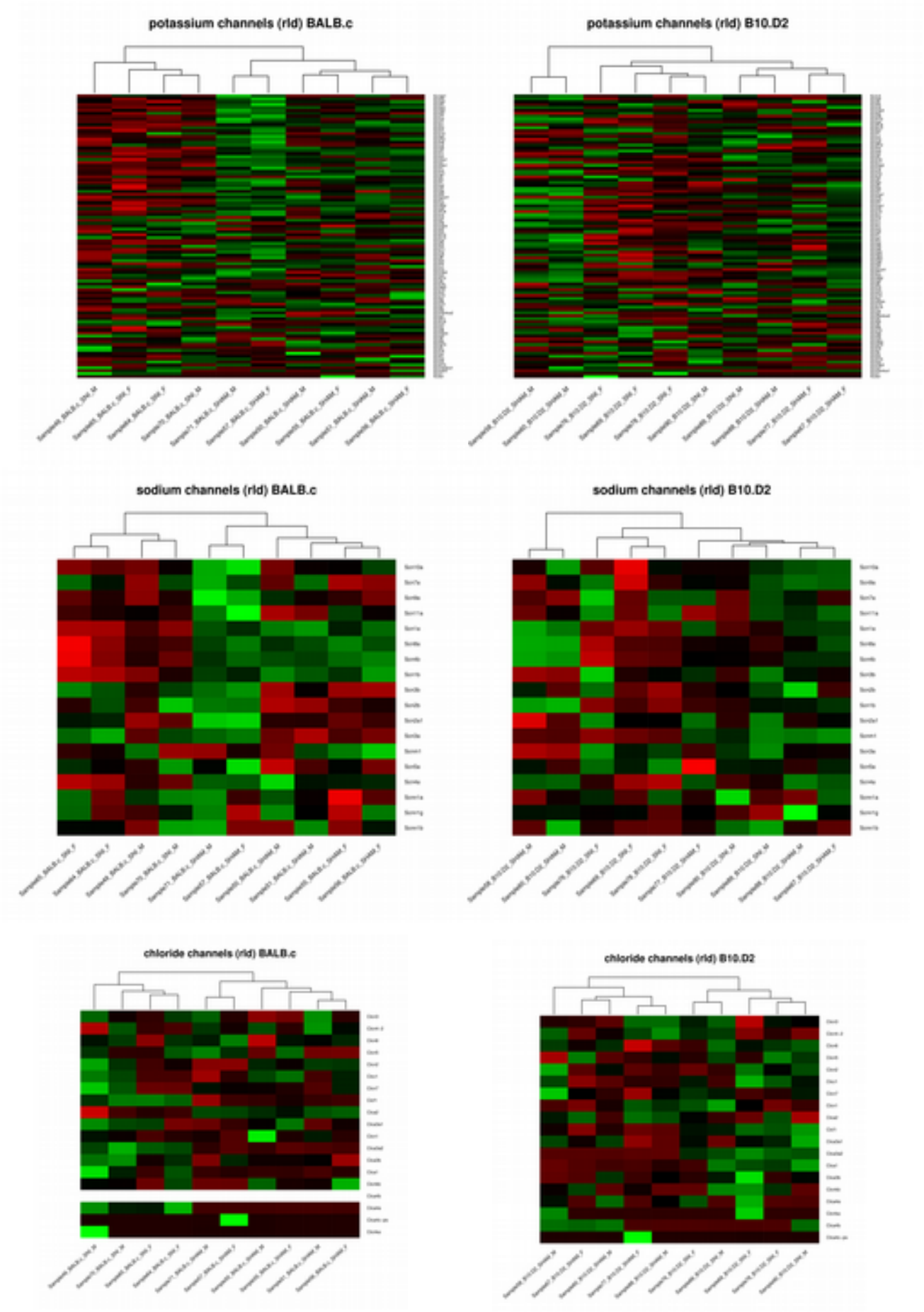


Figure 13: Heatmaps of ion channels for BALB/c strain (left column) and B10.D2 strain (right column). Potassium and sodium channels expression provides optimal separation of samples according to condition for BALB/c strain but not for B10.D2 strain. Chloride channels optimally separated both strains.

Significantly DE genes

Using DESeq2 and the Wald test with the Benjamini–Hochberg correction to control false discovery rate, we assessed the significance of differential expression of genes. We had a total of 16063 genes with nonzero total read counts across conditions. The generalized linear model we fitted to the data gave the following results for each coefficient:

1. ENSEMBL genes significantly differentially expressed (DE) with an adjusted p.value < 0.05 between Male and Female mice:

   LFC > 0 (up) : 165, 0.47%

   LFC < 0 (down) : 8, 0.023%

   outliers : 54, 0.15%

2. ENSEMBL genes significantly DE with an adjusted p.value < 0.05 SNI vs sham mice for the BALB/c (high pain) strain:

   LFC > 0 (up) : 933, 2.7%

   LFC < 0 (down) : 931, 2.7%

   outliers : 54, 0.15%

3. ENSEMBL genes significantly DE with an adjusted p.value < 0.05 SNI vs sham mice for the B10.D2 (low pain) strain:

   LFC > 0 (up) : 578, 1.6%

   LFC < 0 (down) : 687, 2%

   outliers : 75, 0.47%

4. ENSEMBL genes with significantly log fold changes, difference of differences, with an adjusted p.value < 0.05 SNI vs sham for the B10.D2 (low pain) vs SNI vs sham for BALB/c (high pain):

LFC > 0 (up) : 12, 0.034%

LFC < 0 (down) : 24, 0.068%

outliers : 54, 0.15%

These findings reinforced the hypothesis that the injury induced regulation of transcripts at the molecular level is much more prominent for the high pain strain BALB/c. Moreover, we identified a set of 36 genes with significantly different responses / log fold changes in the two strains after the SNI pain surgery.

There is a common core of 846 significantly DE genes between the two strains SNI vs sham. The high pain strain / BALB/c has an additional set of 1018 DE genes, while B10.D2 has an additional set of 419 DE genes, see figure 14.

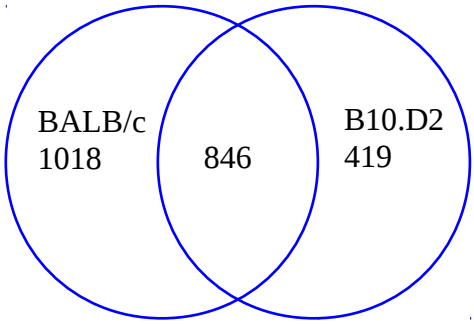


Figure 14: Venn diagram of significantly DE genes in mouse strains

Thus except for a common core of 846 genes the high pain strain extends the repertoire of DE genes by a very significant amount.

Functional Enrichment

By calculating the Gene Ontology functional enrichments for biological process according to several ranking methods (see chapter Methods), we found the top over-represented terms for the DE gene sets presented above. In order to select the top enriched GO-terms we selected the first 20 terms which have the lowest P.value according to the weighted Fisher test and at the same time have a P.value < 0.05 according to the exact Fisher test. For completeness we also included the rank of the GO terms

according to the non-parametric weighted Kolmogorov-Smirnov test. In the following tables we also present the total number of genes annotated with each GO term and the number of significantly DE genes associated with that term in our dataset.

Regarding DE genes SNI vs Sham, in the high pain, BALB/c strain, we observed enrichments of biological processes very relevant to neuropathic pain, table 2. The top 20 terms include terms related to nervous system cell death and regeneration, such as the negative regulation of neuron apoptotic process, positive regulation of synapse assembly, axon guidance, positive regulation of programmed cell death, negative regulation of neuron projection development, positive regulation of cell proliferation and peripheral nervous system development. Moreover there are significantly enriched terms related to signalling and ion channels, like potassium ion transmembrane transport, regulation of ion transmembrane transport, cell adhesion, regulation of calcium ion transport, signal transduction, regulation of potassium transmembrane transport. Terms related to learning and locomotory behaviour have been found to be part of the biological process of pain in functional genomic studies

(Lötsch et al., 2013) and are generally well established in the literature and intuitively accepted. These findings are consistent with the pattern identified from the analysis of the gene expression of ion channels and show that potassium channels are particularly important for the high pain phenotype observed in the BALB/c strain. They also show that a set of processes ranging from signalling, development of axons and neuron cells, behaviour and higher cognitive processes like learning, are essential in establishing pain after peripheral neuropathy.

GO.ID	Term	Significant	Expected	Rank.in.weightKS	P.value classicFisher	P.value weightFisher
GO:0043524	negative regulation of neuron apoptotic process	34	15.7	28	9.197727535066E-06	0.0000092
GO:0071805	potassium ion transmembrane transport	37	15.7	3	4.147362272456E-07	0.0000103
GO:0051965	positive regulation of synapse assembly	20	7.11	8	1.12743454507609E-05	0.0000113
GO:0007411	axon guidance	34	15.92	29	0.000012772	0.0000131
GO:0008360	regulation of cell shape	29	13.32	13	3.71215580856528E-05	0.0000371
GO:0032060	bleb assembly	7	1.24	36	5.05165264741591E-05	0.0000505
GO:0034765	regulation of ion transmembrane transport	65	33.65	6	1.104676982964E-07	0.0000899
GO:0008306	associative learning	18	8.36	127	0.0012	0.0000925
GO:0046661	male sex differentiation	17	10.39	18	0.02747	0.0000986
GO:0007155	cell adhesion	177	119.35	50	2.32207945015E-08	0.00014
GO:0051924	regulation of calcium ion transport	37	20.44	151	0.00024	0.00018
GO:0045109	intermediate filament organization	7	1.47	74	0.00021	0.00021
GO:0007165	signal transduction	516	403.9	12	1.99006995E-11	0.00028
GO:1901379	regulation of potassium ion transmembrane transport	13	5.98	72	0.00511	0.00031
GO:0008344	adult locomotory behavior	24	9.6	194	0.000014836	0.00059
GO:0043068	positive regulation of programmed cell death	72	55.89	124	0.0142	0.00071
GO:0002495	antigen processing and presentation of peptide antigen via MHC class II	5	2.03	17	0.04456	0.00074
GO:0010977	negative regulation of neuron projection development	23	11.86	9	0.00129	0.00086
GO:0008284	positive regulation of cell proliferation	104	74.52	79	0.00024	0.00094
GO:0007422	peripheral nervous system development	13	6.32	92	0.00834	0.0011

Table 2: Top 20 enriched GO terms for biological process in SNI vs Sham for BALB/c strain. The columns hold the GO ID, term description, nr of annotated genes for the term, nr of significantly DE genes for the term, expected nr of significant DE genes, rank according to p.values from wight KS test, p.value for fisher exact test, p.value for weight fisher test

GO.ID	Term	Significant	Expected	Rank.in.weightKS	P.value classicFisher	P.value weightFisher
GO:0043524	negative regulation of neuron apoptotic process	25	10.45	39	3.65475727E-05	0.000037
GO:0048485	sympathetic nervous system development	7	1.28	19	0.00013	0.00013
GO:0019228	neuronal action potential	10	2.56	33	0.00014	0.00014
GO:0007613	memory	21	7.52	270	1.36898081E-05	0.00019
GO:0045109	intermediate filament organization	6	0.98	50	0.00019	0.00019
GO:0044406	adhesion of symbiont to host	6	1.05	5	0.00032	0.00032
GO:0071805	potassium ion transmembrane transport	22	10.45	24	0.00068	0.00044
GO:0021604	cranial nerve structural organization	5	0.83	238	0.00075	0.00075
GO:0051965	positive regulation of synapse assembly	13	4.81	34	0.00081	0.00081
GO:0007155	cell adhesion	111	80.74	293	0.00031	0.00104
GO:0008306	associative learning	15	5.56	259	0.00034	0.00109
GO:0007638	mechanosensory behavior	5	0.9	217	0.0012	0.0012
GO:0007628	adult walking behavior	9	2.78	499	0.00135	0.00135
GO:0030574	collagen catabolic process	6	1.35	142	0.0015	0.0015
GO:0060547	negative regulation of necrotic cell deaath	4	1.05	131	0.01728	0.0016
GO:0050768	negative regulation of neurogenesis	26	16.09	71	0.01032	0.00179
GO:0071417	cellular response to organonitrogen compound	39	23.23	62	0.00098	0.00182
GO:0014059	regulation of dopamine secretion	6	1.73	411	0.00592	0.00184
GO:0010996	response to auditory stimulus	5	0.98	271	0.00184	0.00184
GO:0048385	regulation of retinoic acid receptor signaling pathaway	5	0.98	335	0.00184	0.00184

Table 3: Top 20 enriched GO terms for biological process in SNI vs Sham for B10.D2 strain. The columns hold the GO ID, term description, nr of annotated genes for the term, nr of significantly DE genes for the term, expected nr of significant DE genes, rank according to p.values from weight KS test, p.value for fisher exact test, p.value for weight fisher test

We observed similar enrichment results for the B10.D2 strain SNI vs Sham. The top term was again negative regulation of neuron apoptotic process and several GO terms related to nervous system cell development, like sympathetic nervous system development, positive regulation of synapse assembly, negative regulation of necrotic cell death, negative regulation of neurogenesis. On the other hand, regarding terms related to ion channels, in this strain we only observed enrichment for potassium ion transmembrane support, neuronal action potential and cell adhesion. Interestingly genes DE in this strain were more enriched in terms related to biological processes of higher order related to memory, learning, behaviour and response to stimuli. Also important for the generation of pain sensation, both in brain and the peripheral nervous system is the term associated with to dopamine secretion.

To further analyse differences in the gene expression response between strains after the SNI surgery, we calculated GO enrichments for genes that had significantly different log fold changes between strains. This comparison was very stringent as we did not compare normalised expression values between conditions, but rather log fold changes. Thus we looked for LFCs which were significantly different, so that they cannot be explained by the general trend of DE between SNI vs Sham for the baseline strain. As we only had 26 genes with significantly different responses between strains, the number of genes which were actually associated with the enriched terms were only one or two genes. Nevertheless, we found some terms which may explain the difference in the intensity of pain between strains, table 4.

First, we observed enrichment for genes related to the immune system and hormone secretion. Terms related to T-cells were highly enriched and this confirmed that T-cell activation and infiltration after peripheral nerve injury is a major contributor to hypersensitivity (Costigan et al., 2009). Moreover T-cell’s function is part of the immune component of neuropathic pain which is only present in the central nervous system of adult animals. But more importantly we observed enriched terms for axon

guidance, potassium ion transport and chemokine secrtetion. Being aware that chemokine expression is important in the context of inflammatory response of the immune system after nerve injury (White and Wilson, 2008) and that their upregulation is associated with neural signalling processing, which in turn is crucial for maintaining neuropathic pain (White et al., 2005; White and Wilson, 2008), we can hypothesize that genes that regulate chemokine expression and inflammatory response are important for the differences of pain intensity observed between the two strains. Moreover axon guidance and regeneration of injured neurons is also known to be important in modulation of painful neuropathy after injury in the peripheral or central nervous system. More specifically improved regeneration and axon growth has been found to restore motor behaviour in animals after peripheral neuropathy (Ma et al., 2011). Experimental studies have found that axon regeneration has a strong genetic component (Tedeschi et al., 2016). BALB/c mice after Spinal Cord Injury (SCI), which is a central nervous system pain model, have been found to have significantly less axon regenerative capacity than other strains, including a B10 variant, B10.PL (Basso et al., 2006). Thus induced regeneration of injured neurons, differentially regulated inflammatory response and chemokine secretion might be the driving factors of the less intense mechanical hypersensitivity observed in the B10.D2 strain compared to the BALB/c strain. Moreover in the BALB/c strain 25% of pain genes were called as significant DE with adjusted p.values < 0.05 while in the B10.D2 strain we observed only 15.9%. Thus the observed differences in pain intensity could be also associated with the number of significantly dysregulated pain genes.

GO.ID	Term	Annotated	Significant	Expected	Rank.in.weightKS	P.value classicFisher	P.value weightFisher
GO:0032350	regulation of hormone metabolic process	38	2	0.03	1610	0.00049	0.00049
GO:0045625	regulation of T-helper 1 cell differentiation.	10	1	0.01	572	0.00866	0.00866
GO:0045628	regulation of T-helper 2 cell differentiation	10	1	0.01	1573	0.00866	0.00866
GO:0051798	positive regulation of hair follicle development	10	1	0.01	1380	0.00866	0.00866
GO:0090197	positive regulation of chemokine secretion	10	1	0.01	1889	0.00866	0.00866
GO:2000849	regulation of glucocorticoid secretion	10	1	0.01	1939	0.00866	0.00866
GO:0007411	axon guidance	165	2	0.14	3336	0.00882	0.00882
GO:0046632	alpha-beta T cell differentiation	80	2	0.07	4276	0.00215	0.00896
GO:0002830	positive regulation of type 2 immune response	12	1	0.01	119	0.01038	0.01038
GO:0006590	thyroid hormone generation	12	1	0.01	3814	0.01038	0.01038
GO:0045623	negative regulation of T-helper cell differentiation	12	1	0.01	3158	0.01038	0.01038
GO:0006521	regulation of cellular amino acid metabolic process	13	1	0.01	4482	0.01124	0.01124
GO:0007210	serotonin receptor signaling pathway	14	1	0.01	3945	0.0121	0.0121
GO:0045624	positive regulation of T-helper cell differentiation	14	1	0.01	717	0.0121	0.0121
GO:2000833	positive regulation of steroid hormone secretion	14	1	0.01	4270	0.0121	0.0121
GO:0070633	transepithelial transport	15	1	0.01	4200	0.01296	0.01296
GO:1901018	positive regulation of potassium ion trasport	15	1	0.01	4575	0.01296	0.01296
GO:0007202	activation of phospholipase C activity	16	1	0.01	3163	0.01382	0.01382
GO:0043032	positive regulation of macrophage activation	16	1	0.01	1307	0.01382	0.01382
GO:0043306	positive regulation of mast cell degranulation	16	1	0.01	1710	0.01382	0.01382

Table 4: Top 20 enriched GO terms for biological process in genes with significant different response between strains. The columns hold the GO ID, term description, number of annotated genes for the term, number of significantly DE genes for the term, expected number of significant DE genes, rank according to p.values from weight KS test, p.value for Fisher exact test, p.value for weight Fisher test

We next examined the direction of changes in the expression of the most interesting of the genes that were found to have significantly different (adjusted p.value < 0.05) responses between BALB/c and B10.D2 strains.

Acan, which is associated with calcium ion binding and extracellular matrix has been found to be up-regulated in patients with painful Femoroacetabular Impingement (Chinzei et al., 2016), was significantly up-regulated in BALB/c strain (p.value = 9.975E-06, log fold change = 2.51) while it was slightly down-regulated in B10.D2 strain (log fold change = -0.54).

Duoxa1, which is a p53-regulated neurogenic factor whose expression is induced by overexpression of p53 and intensifies neuronal differentiation (Ostrakhovitch and Semenikhin, 2011), was also significantly up-regulated in BALB/c (p.value = 5.589E-030, log fold change = 3.02) and not DE in B10.D2.

Gal, a neuropeptide involved in nociception which functions as a cellular messenger of the nervous system and is related to neuropathic pain (Mechenthaler, 2008), was significantly more upregulated in the BALB/c (log fold change = 4.4) strain than in the B10.D2 strain (log fold change = 1.8).

Interestingly though, Il4ra which encodes the alpha chain of the interleukin-4 receptor and is found to be implicated in inflammatory macrophage-dependent neuropathic pain (Kiguchi et al., 2015), was stable in B10.D2 strain but significantly up-regulated in BALB/c strain after the SNI surgery (p.value = 1.18E-08, log fold change = 0.7). Even though Il4ra is known to induce M2 macrophages which reduce neuroinflammation (Casella et al., 2016). Il4ra has also been found to reduce neuropathic pain in mice after sciatic nerve injury, (Kiguchi et al., 2015) .

After SNI surgery Chl1 and Gap43, which are related to synaptic plasticity and axon guidance, and are known to be crucial for neuronal regeneration after nerve injury (Cheng et al., 2013; Yamanaka et al., 2011) , were significantly more up-regulated in BALB.c (high pain strain) than in

B10.D2 strain. Increased maladaptive plasticity of the peripheral and central nervous system is indeed a characteristic of both neuropathic and inflammatory pain (Michael Costigan et al., 2009).

Identification of LncRNAs

After analysing the expression of the known genes we proceeded to identify novel LncRNAs and analyse their differential expression. We used the pipeline described in chapter methods to identify 4970 predicted LncRNAs with non-zero counts. Next we assessed the coding potential of these putative LncRNAs and found 936 with a coding potential score, as calculated by CPC, of more than 1. Thus we discarded 936 identified transcripts as coding and we ended up with 4034 predicted models of putative LncRNAs.

Mouse is a very well annotated organism and since we used the latest mouse genome (mm10) we observed 402 already annotated or predicted LncRNAs by the ENSEMBL/GENCODE consortium pipeline, expressed in our dataset. We then compared and looked for overlaps between these ENSEMBL LncRNAs and the ones predicted from RNA-seq using our customised pipeline. Using our pipeline, we were able to identify 317 (78.8%) out of these 402 ENSEMBL LncRNAs with more than 50% exonic sequence overlap. A plot of the log mean counts of identified vs missed LncRNAs, figure 15, shows that the average log 2 counts for the missed LncRNAs are significantly lower than for those identified by our pipeline. As our aim is not a complete reconstruction of the transcriptome, but rather finding novel LncRNAs, which may be functionally important in certain biological processes because they are significantly DE, we identified LncRNAs which are sufficiently and consistently expressed in order for their expression pattern to be further analysed. These results indicate that by controlling possible false positives and sequencing artefacts, our pipeline identified a stringent set of putative LncRNAs that could have functional importance in specific biological conditions.

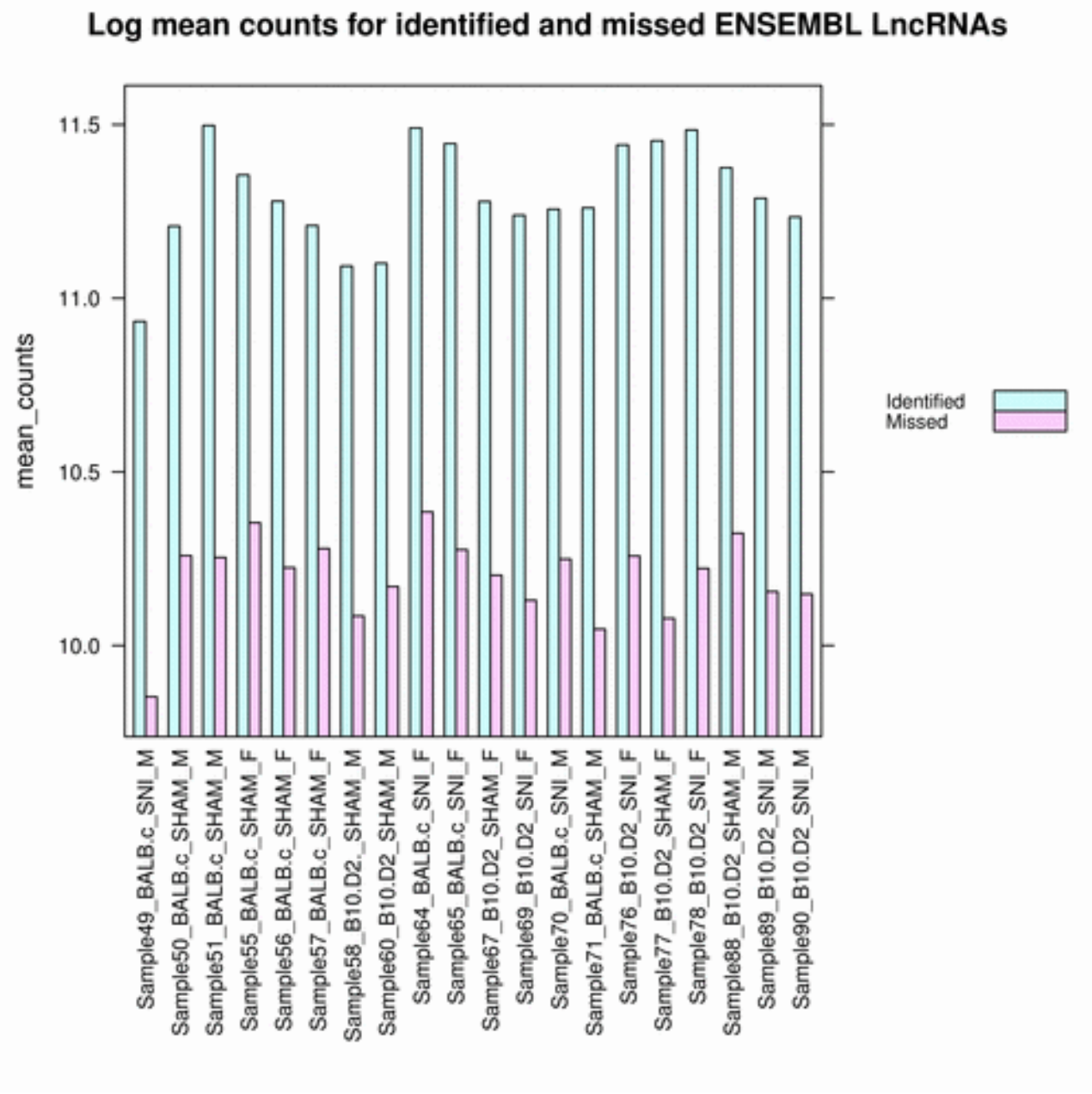


Figure 15: Log mean counts of Identified and Un-identified ENSEMBL LncRNAs in the mouse RNA-seq dataset.

Moreover, before assessing coding potential, we examined the predicted LncRNAs in the context of their exon numbers. GENCODE has found that most of annotated LncRNAs have two exons (Harrow et al., 2012). Our predictions have a distribution of exon number which is very similar to GENCODE's findings. The distribution of exons is heavily skewed to the left and biased towards bi-exonic LncRNAs, then tails off

heavily after 5 exons, with near zero predictions of more than 8 exons, figure 16.

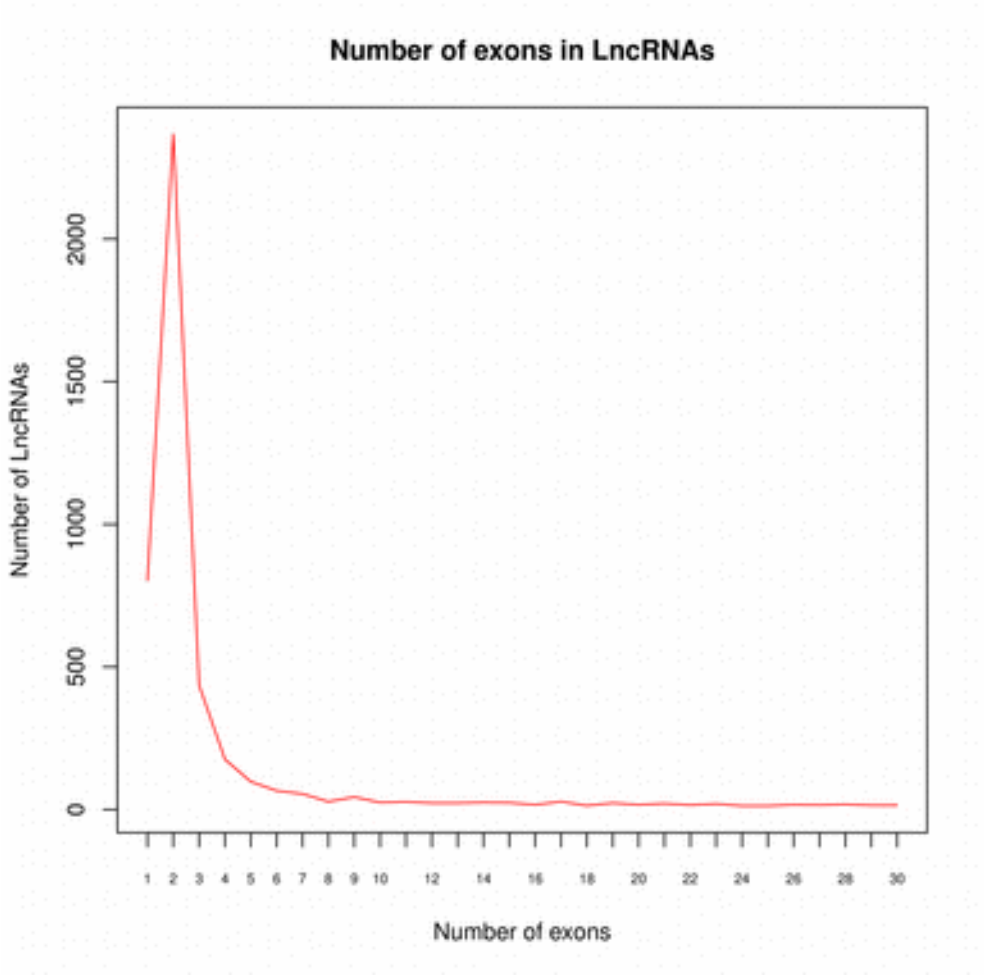
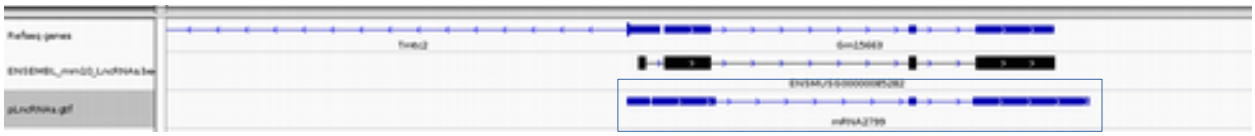


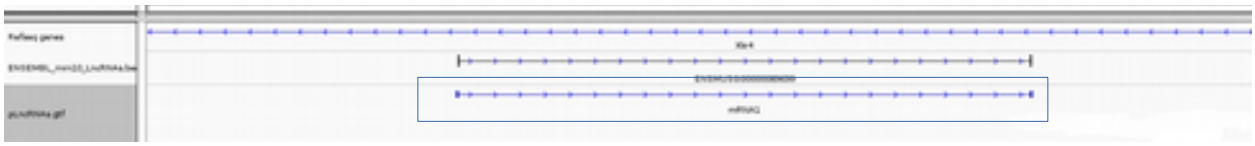
Figure 16: Distribution of exon number in the predicted LncRNAs. Most of the transcripts are predicted to have 2 exons.

Some examples of annotated LncRNAs identified by our pipeline are in figure 17.

1)

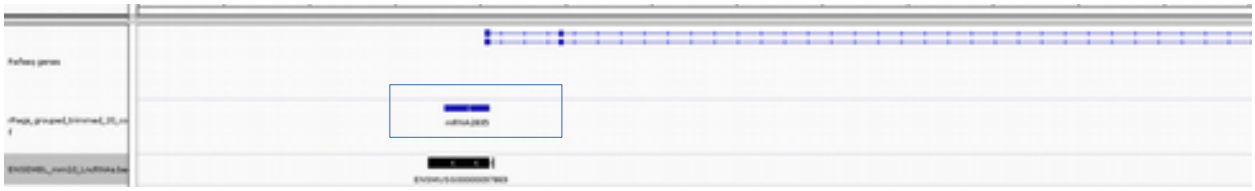


2)



3)

4)



5)

Figure 17: Annotated LncRNAs as predicted in our dataset. For each IGV genome browser screenshot: in the first track named "RefSeq genes" is the RefSeq annotation, in the second track named "ENSEMBL_mm10_LncRNAs" is the ENSEMBL annotation, in the third track is our predicted gene models of LncRNAs (in the blue box). From top to bottom:

1. *ENSMUSG00085282 ENSEMBL LncRNA. This is a complex transcript with 4 exons of different lengths. We should note the significant differences between ENSEMBL and RefSeq annotations. Our predicted LncRNA matches the ENSEMBL annotation in exon number and structure.*
2. *ENSMUSG00000089699 ENSEMBL LncRNA. Refseq does not include this LncRNA in its mm10 annotation. We predicted the exact gene model given by ENSEMBL. This is bi-exonic LncRNAs with two relatively small exons.*
3. *NR_040391 RefSeq LncRNA. ENSEMBL pipeline has not predicted this LncRNA and it does not include it in its mm10 annotation. We predicted a gene model, with not good agreement with RefSeq with two exon less than the RefSeq annotation.*
4. *ENSMUSG00000097869 ENSEMBL LncRNA. This is a bi-exonic antisense LncRNA with a very small exon and a larger one. We predicted both exons. RefSeq annotation does not include this LncRNA at all.*
5. *ENSMUSG00000087413 ENSEMBL LncRNA. This is again a bi-exonic LncRNA. We predicted both exons. RefSeq annotation does not include this LncRNA at all.*

Expression of predicted LncRNAs in mouse DRG

After assigning reads to the identified LncRNAs using HTSeq (Anders et al., 2015) and the *Intersection Not Empty* strategy, we examined their expression strength relative to protein coding genes. Predicted LncRNAs had, as expected, significantly lower median expression than protein coding genes. Median values of LncRNAs were 10 times lower than those for ENSEMBL protein coding genes. This pattern was consistent in all samples, figure 18. As we have deeper sequencing for the mouse samples, i.e. more reads per sample, and thus higher dynamic range we observed much higher difference between the median expression of ENSEMBL protein coding genes and LncRNAs in mouse than in rat (10 fold compared to 5 fold). In mouse we had 70-80 million reads per sample vs 50 million reads per sample in rat, but the difference was also due to the much better mapping of reads, i.e. the actual number of reads used for downstream analysis, in mouse than in rat.

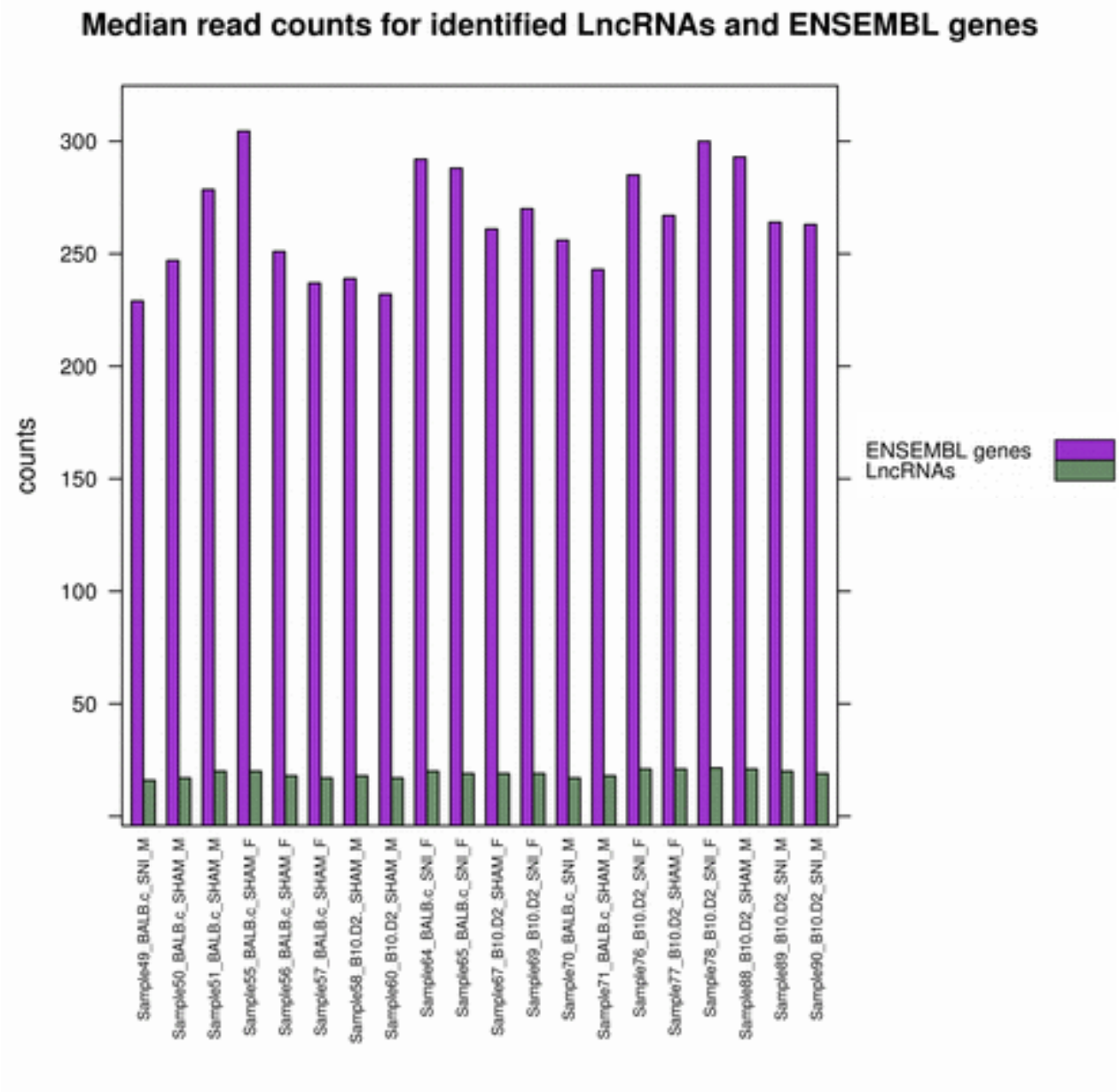


Figure 18: Median read counts for predicted LncRNAs (green) compared to median read counts for ENSEMBL protein coding genes (violet). In general LncRNAs were expressed 10 times lower than genes.

Differential Expression of LncRNAs

We subsequently analysed differential expression of LncRNAs using DESeq2. First we assessed the quality of our data by inspecting how the expression of these novel LncRNAs separated our samples, see figure 19. This would indicate whether they carry a biological signal relevant to neuropathic pain. Then we assessed the consistency of the expression of LncRNAs in our dataset by calculating the Cook's distance for each sample, i.e. the change of the coefficient of a linear model fitted to the LncRNAs' expression if we remove the respective sample and refit the model. A consistent Cook's distances plot would indicate that we did not have spurious expression spikes, genomic contamination or highly inconsistent expression of novel LncRNAs, figure 20. Regarding maximum Cook's distance which can reveal whether there are lot of outlying LncRNAs, we observed a very similar distribution to that of ENSEMBL genes, with very few outliers and generally small distances. The only expected difference is that the values for LncRNAs were 10 times smaller than the values for ENSEMBL genes, following their fewer numbers and lower expression strength, figure 21.

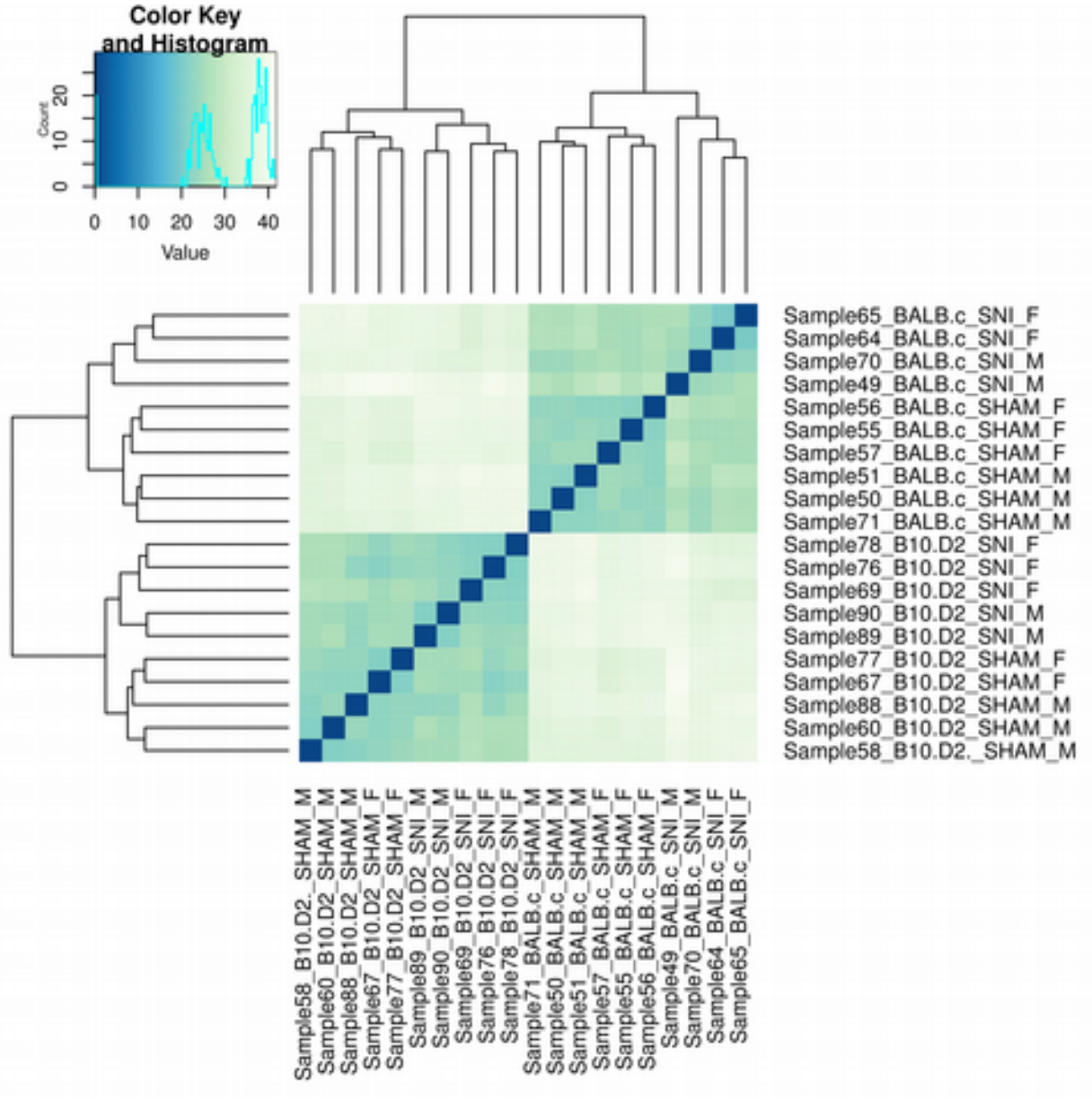


Figure 19: Clustering of samples according to the expression of predicted LncRNAs. Samples are clustered in the same fashion as for known genes, first by strain and then by condition and within these two by gender.

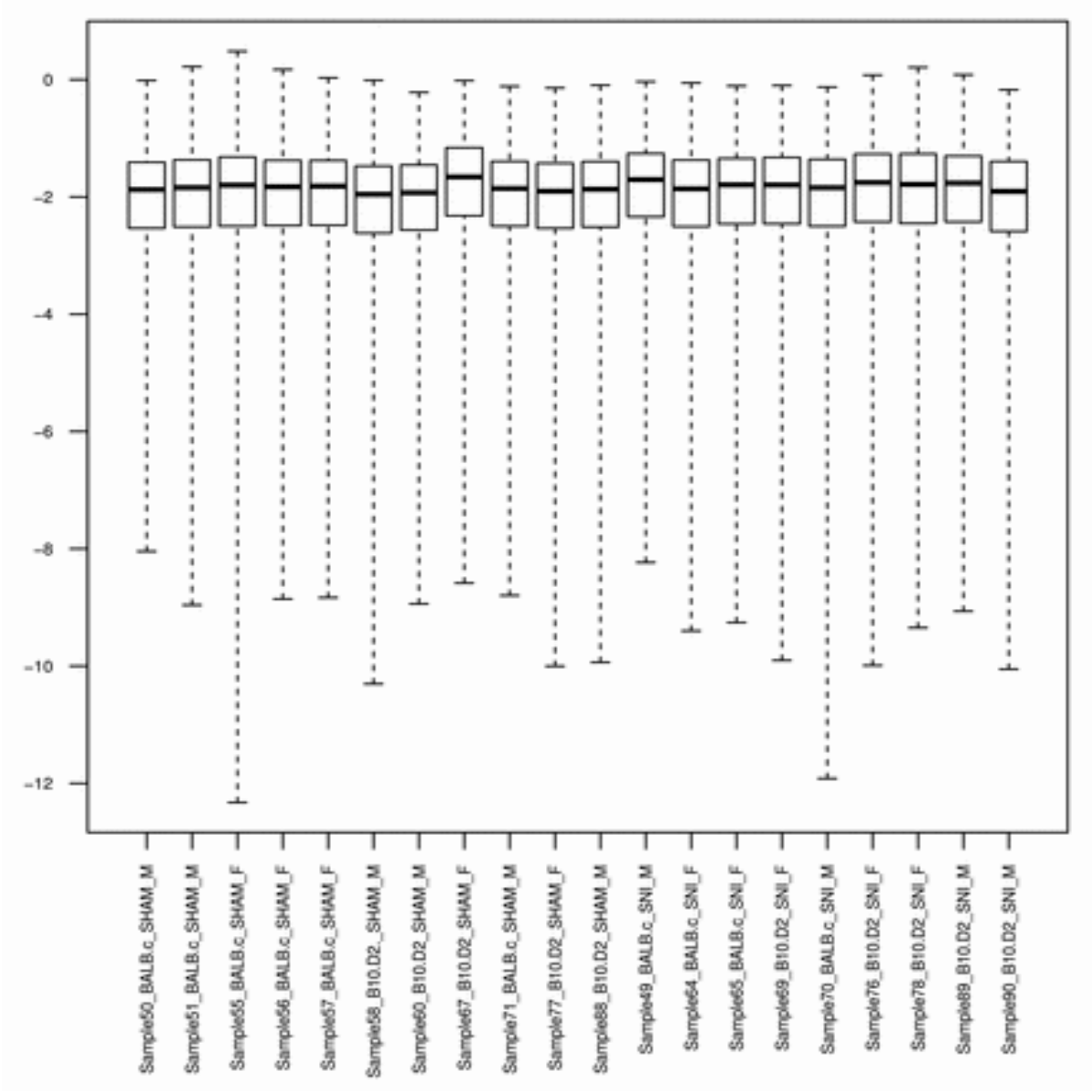
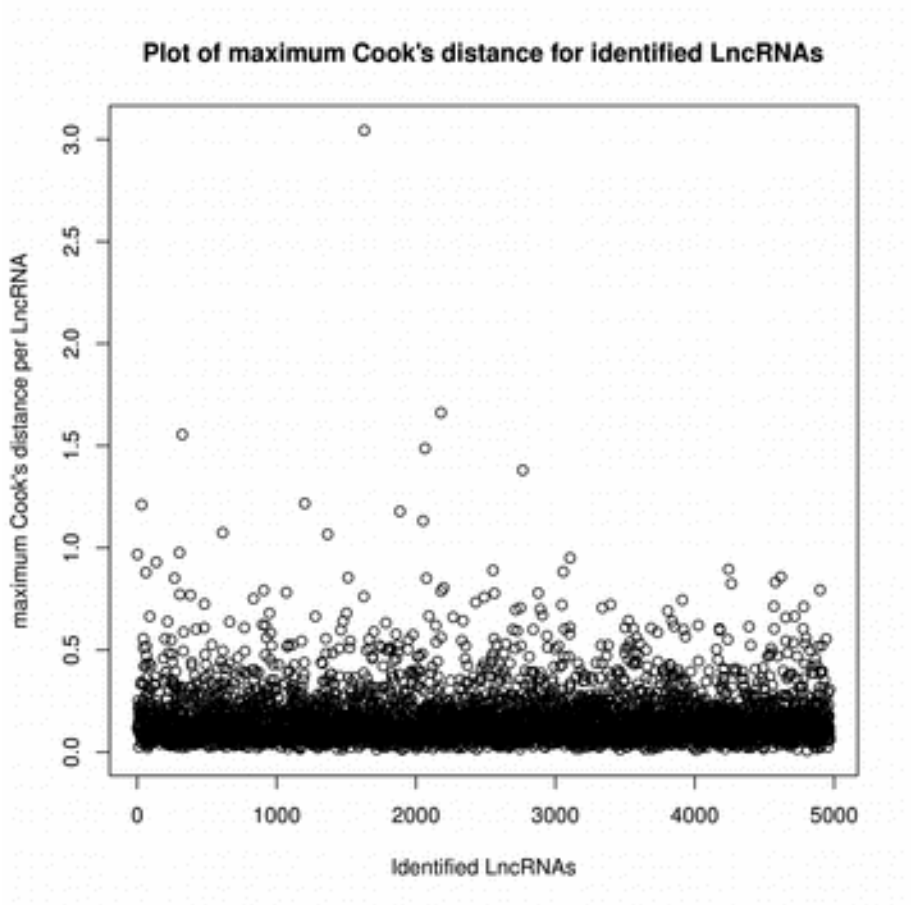


Figure 20: Log10 Cook's distance of the expression of novel LncRNAs for each sample in our mouse RNA-seq dataset. Very similar Cook's distance in terms of median value and interquantile range for each sample, indicated very consistent expression of novel LncRNAs.



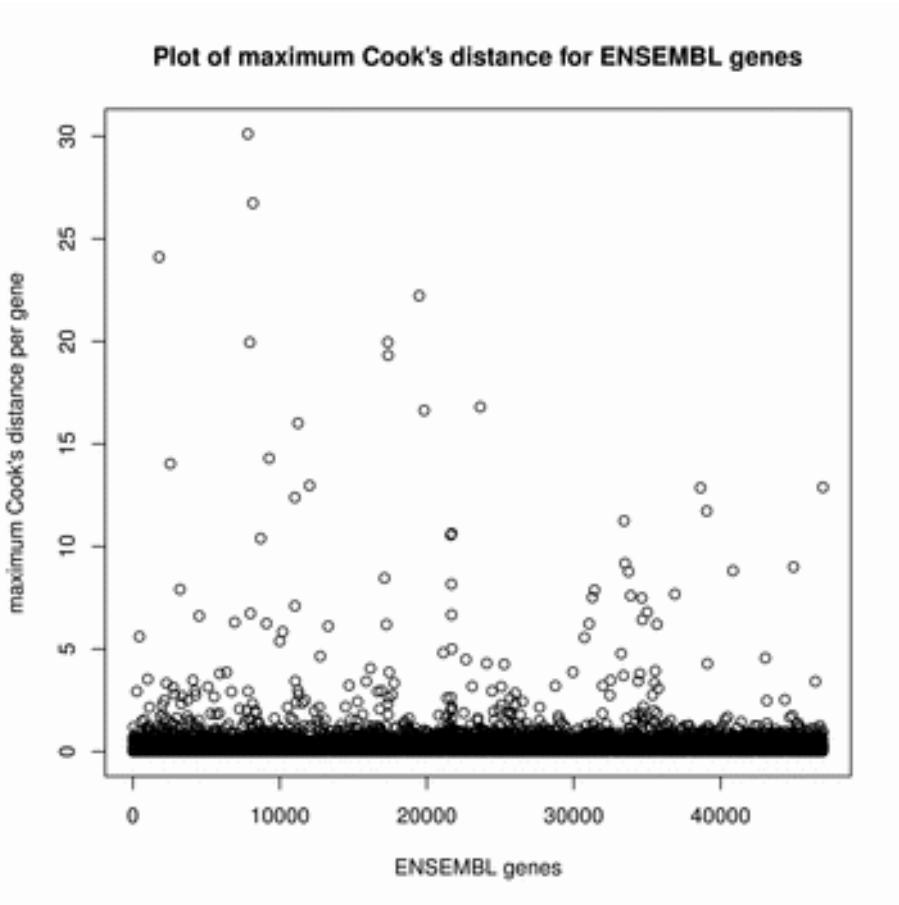


Figure 21: Maximum Cook's distance of LncRNAs (left) and ENSEMBL genes (right)

After successful quality control we performed DE analysis. Using DESeq2 and the Wald test with the Benjamini–Hochberg correction to control false discovery rate and adjust p.values, we assessed the significance of differential expression of predicted LncRNAs. We had a total of 4970 LncRNAs with nonzero read counts across conditions. By fitting the generalized linear model we obtained the following results for each coefficient:

1. Predicted LncRNAs significantly differentially expressed (DE) with an adjusted p.value < 0.1 between Male and Female mice:

   LFC > 0 (up) : 2, 0.04%

   LFC < 0 (down) : 1, 0.02%

   outliers : 97, 2%

2. Predicted LncRNAs significantly DE with an adjusted p.value < 0.05 SNI vs sham mice for the BALB/c (high pain) strain:

   LFC > 0 (up) : 101, 2%

   LFC < 0 (down) : 70, 1.4%

   outliers : 0, 0%

3. Predicted LncRNAs significantly DE with an adjusted p.value < 0.05 SNI vs sham mice for the B10.D2 (low pain) strain:

   LFC > 0 (up) : 60, 1.2%

   LFC < 0 (down) : 64, 1.3%

   outliers : 97, 2%

4. Predicted LncRNAs with significantly different log fold changes (difference of differences) with an adjusted p.value < 0.05 SNI vs

sham for the B10.D2 (low pain) vs SNI vs sham for BALB/c (high pain):

LFC > 0 (up) : 0, 0%

LFC < 0 (down) : 6, 0.234%

outliers : 97, 2%

Although there were fewer predicted LncRNAs found to be significantly DE than ENSEMBL genes, in terms of percentages the numbers are very similar and the above findings are very similar to those obtained for known ENSEMBL genes in terms of the extent of significant dysregulation across strains and conditions. This data reinforced the observation that the pain response at the molecular level is more prominent for the high pain strain BALB/c. These observations suggested that both, for known genes and predicted LncRNAs, there is more significant dysregulation in the high pain strain BALB/c, than in the low pain strain B10.D2. We should also note that the median counts across conditions were significantly lower than that of genes, a finding consistent with the literature for LncRNAs (see chapter Introduction, section Long non-coding RNAs).

As we were interested to study all novel gene models / transcribed loci predicted from our pipeline, we calculated DE for all of them and then we discarded the ones that showed positive coding potential. Regarding coding potential scores we had an enrichment of transcripts with positive coding potential amongst the predicted LncRNAs which were significantly DE. This is reasonable, as gene models found to be protein coding are expected to have higher expression strength, thus higher read counts, which leads to a more confident estimation of the actual log fold changes and of differential expression. Thus, it is reasonable to have relatively more predicted protein coding transcripts in the set of transcripts with lower p.values, figure 22. Out of the 6 transcripts initially identified as having significantly different responses to SNI between strains, 2 had positive coding potential. Thus only the remaining 4 can be considered LncRNAs.

On the other hand, it is evident from the plot in figure 21 that the two other comparisons are much less affected from excluding the predicted transcripts with high coding potential.

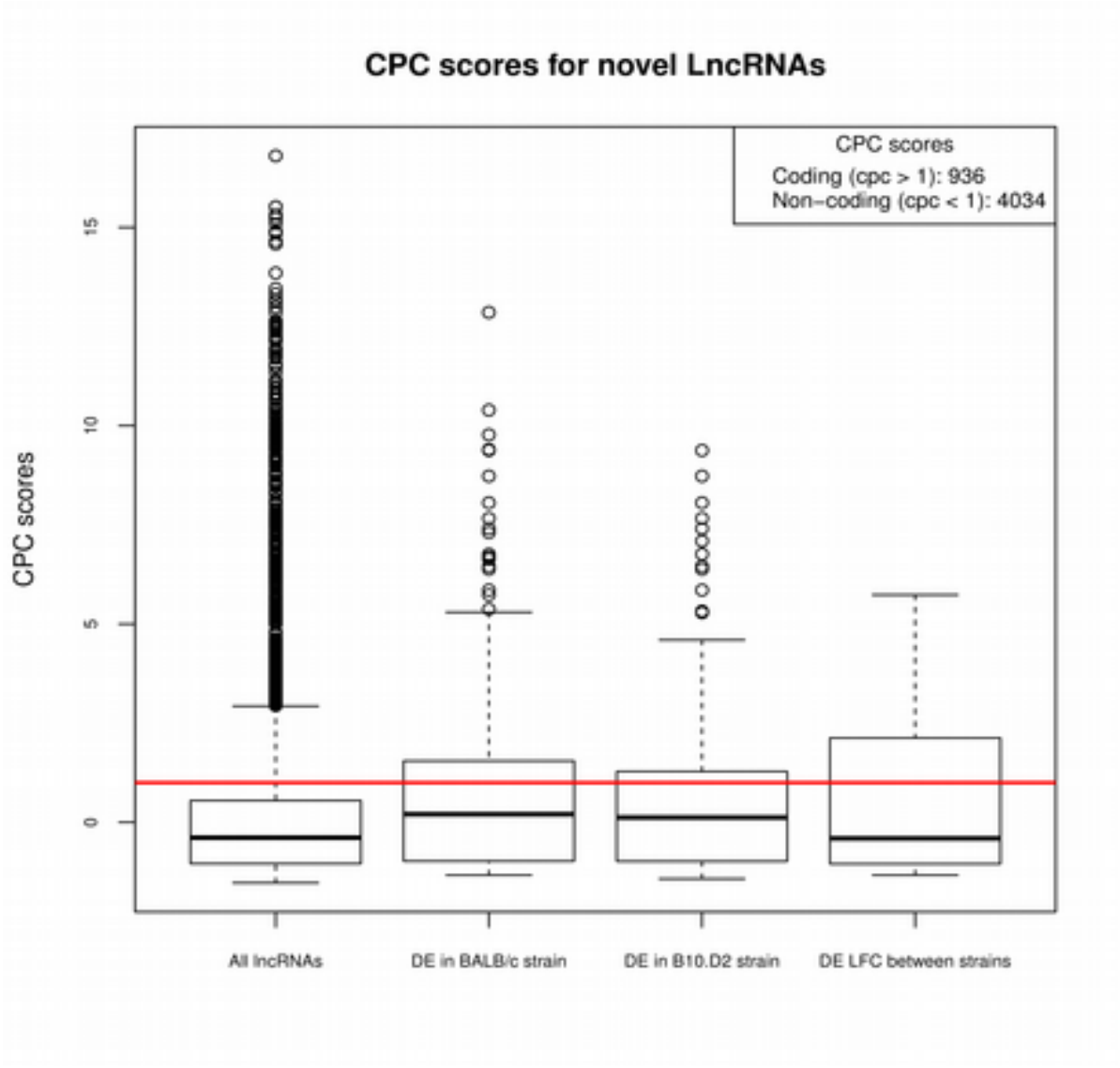


Figure 22: Distribution of CPC scores for all predicted LncRNAs and for those significantly DE (adj. p.value < 0.05). The red line represents the threshold of CPC score = 1. Above it transcripts are considered protein coding.

Antisense LncRNAs and pain-related protein coding genes

In order to infer the functional role of some of the novel LncRNAs identified we examined them alongside their genomic context. Thus we selected all the novel LncRNAs which overlap any of the genomic region of protein coding genes on the opposite strand. In total we identified 2415 expressed antisense LncRNAs on the opposite strand of 3568 protein coding genes. We did not limited our analysis only to LncRNAs antisense of pain genes, but we selected and further discussed above the LncRNAs identified significantly DE and antisense of significantly DE genes.

Not many of the identified antisense LncRNAs were significantly DE. In BALB/c strain SNI vs sham we had 8 LncRNAs significantly DE (adjusted p.value < 0.05) antisense of significantly DE (adjusted p.value < 0.05) protein coding genes. More specifically we identified LncRNAs antisense of Tpd52l1, Nalcn, Scn1a, Tshz2, Ttc39a, Arhgap35, Gm19424 and Ralyl. The LncRNAs antisense of Nalcn and Tshz2 were anti-correlated to the protein coding gene, i.e. in both cases the antisense LncRNAs is significantly up-regulated while the sense gene is significantly down-regulated.

The antisense LncRNA (log fold change = 3.5, p.value = 1.33E-22, cpc score = -0.44) of Nalcn (log fold change = -0.23, p.value = 0.004) was not found to be significantly DE in B10.D2 strain and neither was Nalcn gene. Thus we identified a couple of protein coding gene and antisense LncRNA, which had opposite expression patterns and were both significantly DE only in high pain strain BALB/c. Nalcn is an important gene in neuropathic pain, it is a sodium-leak channel, which regulates neuronal excitability (Lu et al., 2010).

Tshz2 (log fold change = -0.268, p.value = 0.009), a gene which is expressed in the nervous system and regulates developmental processes, and its antisense LncRNA (log fold change = 2.48, p.value = 3.98E-7, cpc score = -1.247) are another pair of protein coding gene and antisense LncRNA that were not DE in the low pain strain B10.D2.

In B10.D2 strain we had 8 significantly DE LncRNAs antisense of significantly dysregulated genes. In this strain we found a significantly DE LncRNA (cpc score = -0.852) antisense of sodium channel Scn1a. Both the protein coding gene and the LncRNA were significantly down-regulated after SNI surgery. The same is true for Nrp2 gene, an important gene associated with axon guidance and innervation of inner organs (Maden et al., 2012).

By looking at transcription antisense of pain genes downloaded from the PainGenes database (Lacroix-Fralish et al., 2007) we identified 73 expressed transcripts antisense of pain genes. Out of these 60 were putative LncRNAs (cpc score < 1). One of these is the antisense LncRNA of the Scn9a gene. This particular antisense transcript of Scn9a has been recently identified and published (Koenig et al., 2015), although it is not included on the ENSEMBL annotation. In Koenig et al., this transcript was not found to be DE after SNI surgery and this is also the case for the samples in our dataset. However, this specific antisense LncRNA was found to be significantly DE in rat after SNT surgery. All the LncRNAs that were found antisense of pain genes, their cpc score and their DE analysis for SNI vs Sham for the BALB/c strain are shown in Appendix 4.

In B10.D2 strain we found 72 expressed transcripts antisense of pain genes, and 60 of them are putative LncRNAs as they do not show significant coding potential. All LncRNAs antisense of pain genes together with their coding potential and DE analysis for SNI vs Sham in B10.D2 strain are in Appendix 4.

All LncRNAs antisense of pain genes expressed in B10.D2 mouse DRG were also expressed in BALB/c mouse DRG. This consistency in expression reinforces the hypothesis, that even though most of them were lowly expressed and they are not found to be significantly DE after SNI surgery, they are actual antisense LncRNAs which may have a functional role regulating the protein coding gene on the opposite strand.

Intergenic LncRNAs and pain genes

Long Intergenic non-coding RNAs (LincRNAs) can also regulate protein coding gene expression *in-cis* or *in-trans*. Sometimes the product of transcription, the LincRNA *per se* is important for regulating gene expression, whereas in other cases the act of transcription itself induces gene expression. For more details see chapter Introduction, section Long non-coding RNAs. We identified 2096 LincRNAs in total with no coding potential. 101 were found to be significantly DE (adj. p.value < 0.05) in SNI vs Sham in BALB/c strain and 79 in B10.D2 strain. 43 of them had one pain gene as their closest genomic feature. We selected and discuss below all LincRNAs significantly DE, in close proximity and highly correlated with significantly DE protein coding genes.

53 LincRNAs were significantly DE (adj. p.value < 0.05) and adjacent to significantly DE ENSEMBL genes in BALB/c SNI vs sham and 47 in B10.D2 SNI vs sham. In terms of distance between LincRNAs and ENSEMBL genes, half of them are less than 32.5Kb away from genes and 25% of them less than 7Kb, with a few outliers being in more remote genomic regions with no adjacent genes. Moreover when we considered only the pairs of LincRNA and adjacent genes which were both significantly DE in both strains, there were no distant LincRNAs, figure 23.

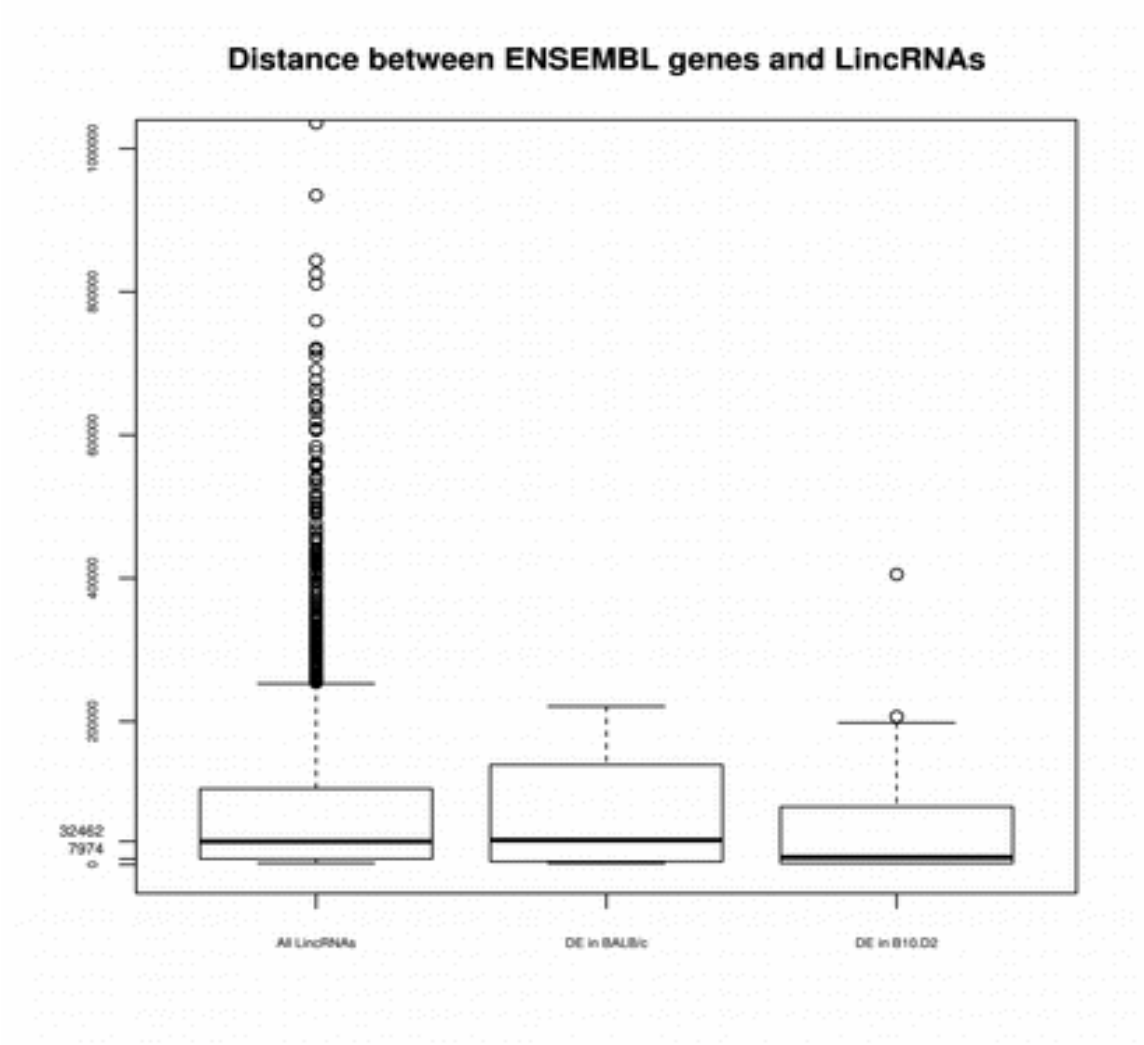


Figure 23: Distance between ENSEMBL genes and LincRNAs. Most of them are in close and moderate genomic proximity.

In order to find putative LincRNAs likely to be *in-cis* regulators of the expression of pain genes, we looked for highly correlated pairs of protein coding pain genes and adjacent LincRNAs.

Since genomic features of close proximity usually have correlated expression (Thygesen and Zwinderman, 2005) and the LincRNAs follow the same pattern (Ulitsky and Bartel, 2013), we established a correlation threshold using a randomization/permutation approach described in Methods.

We considered all LincRNAs with significant correlation estimates (p.value < 0.05) and a correlation coefficient of normalised read counts across samples higher than 0.67 to be highly correlated. In order to account for the difference in scales and also in order to moderate fold changes for LincRNAs with low counts, we used the regularised log transformation of read counts for both ENSEMBL genes and LincRNAs.

lincRNA	ENSEMBL_id	Distance	symbol	cpc	correlation	cor_pvalue
chr4:132107492-132108725(-)	ENSMUSG00000050511	-2001	Oprd1	-1.02247	0.9366318108	1.24159438286142E-09
chr13:54233180-54236274(+)	ENSMUSG00000034987	-10748	Hrh2	0.140287	0.8928979827	1.18692713879653E-07
chr15:72505532-72506966(-)	ENSMUSG00000036760	-5153	Kcnk9	-0.981673	0.9375715301	1.08915854113434E-09
chr19:22992291-22995602(+)	ENSMUSG00000052387	-2407	Trpm3	-1.22812	0.8618264664	1.04443712678481E-06

Table 7: LincRNAs with highly correlated expression to pain genes

Using this approach we found 4 LincRNAs with significantly correlated expression to their adjacent pain gene. All LincRNAs had positive correlation to their closest pain gene. These LincRNAs are adjacent to pain genes Oprd1, Hrh2, Trpm3 and Kcnk9, table 7. They are all upstream of their adjacent pain gene in very close genomic proximity.

The only pair of LincRNA and adjacent gene that were both significantly DE in both strains is the pair of Oprd1 gene and chr4:132107492-132108725(-) LincRNA, figure 25. Oprd1 is an important pain gene, it reduces calcium ion currents and induces potassium ion conductance. It is an opioids receptor and mediates opioids analgesia. As we recently demonstrated, contribution of endogenous opioids leads to congenital insensitivity to pain in humans and mice (Minett et al., 2015). Furthermore, it has been found that genetic differences in heat sensitivity might be related to sex-specific mediation of opioid analgesia by the Oprd1 gene (Mogil et al., 1997). Indeed in our dataset Oprd1 is more downregulated in the high pain strain and the same is also true for the adjacent LincRNA.

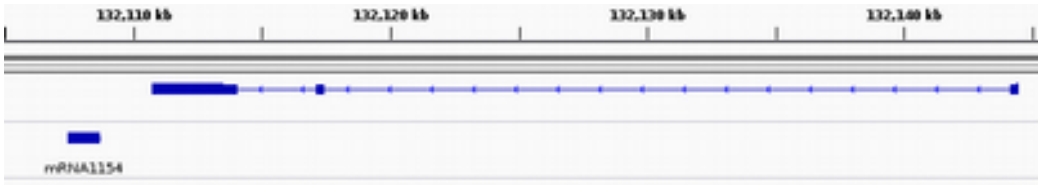


Figure 25: Oprd1 pain gene and chr4:132107492-132108725(-) adjacent LincRNA in the genome browser

Comparing the mouse results to rat under the SNT pain model

Examining the above results in the context of the results obtained in rats that underwent the SNT pain model we observed that the amount of significantly dysregulated genes is much higher in rat than in mouse. This could be due to the higher within samples dispersion we observed in mouse samples. But more interestingly the high pain mouse strain BALB/c is much

more similar to rat than the low pain mouse strain B10.D2, both in the amount of significantly dysregulated genes and their direction of fold change, figure 26. 434 ENSEMBL genes are commonly up-regulated in BALB/c strain and rat and 393 are commonly down-regulated. The respective numbers for the B10.D2 strain are 288 and 253.

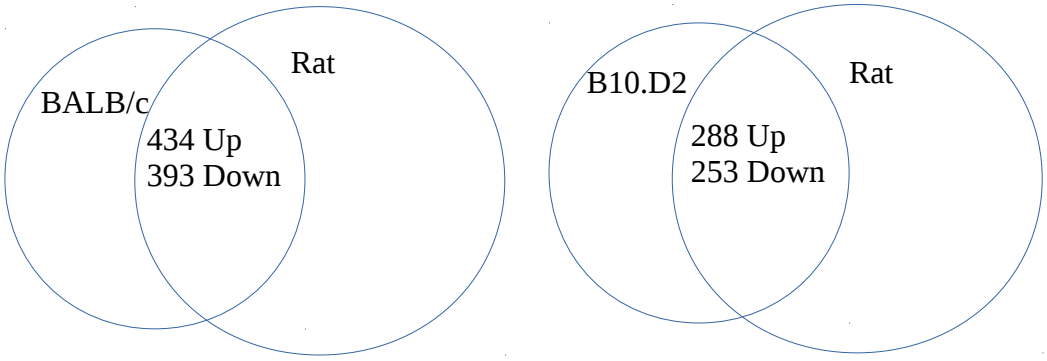


Figure 26: Venn diagram of DE genes in B10.D2, BALB/c mice and rat

We subsequently identified the significantly over-represented gene ontology (GO) terms of biological processes in these subsets of commonly DE genes, table 8 and 9. For this particular enrichment we used as the gene universe all significantly DE genes in either strain, then with a binary function we selected only the ones the were significantly DE in both strains. In the top 10 enriched GO terms we again found enrichment for axon guidance, regulation of apoptosis and cell proliferation, terms related to immune system; signal transduction, cell matrix adhesion and extracellular matrix organisation and only in BALB/c and rat, G-protein coupled receptor signalling. Thus this subset of common terms provides a summarisation at the level of biological processes of the common core of genes dysregulated in both pain models (SNI and SNT) and species (mouse and rat).

GO.ID	Term Description	Annotated	Significant	Expected	Rank in weightKS	p.value classicFisher	p.value weightFisher
GO:0032060	bleb assembly	11	7	0.22	5260	4.0e-10	4.0e-10
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	998	48	20.15	30	2.4e-08	4.8e-07
GO:0035590	purinergic nucleotide receptor signaling pathway	10	5	0.2	5299	7.6e-07	7.6e-07
GO:0043029	T cell homeostasis	44	8	0.89	5275	2.4e-06	2.4e-06
GO:0007165	signal transduction	5302	191	107.03	2278	2.4e-19	7.8e-06
GO:0007160	cell-matrix adhesion	171	19	3.45	5285	1.9e-09	8.5e-06
GO:0043524	negative regulation of neuron apoptotic process	152	13	3.07	245	1.3e-05	1.3e-05
GO:0030198	extracellular matrix organization	201	15	4.06	3898	1.5e-05	1.7e-05
GO:0007411	axon guidance	168	13	3.39	114	3.8e-05	2.8e-05
GO:0035588	G-protein coupled purinergic receptor sinalling	16	4	0.32	5307	0.00025	3.2e-05

Table 8: GO enrichment for DE genes with the same direction in both BALB/c mouse strain and rat

GO enrichment for DE genes with the same direction in both B10.D2 mouse strain and rat

GO.ID	Term Description	Annotated	Significant	Expected	Rank in weightKS	classicFisher	weightFisher	weightKS	elimKS
GO:0032825	positive regulation of natural killer cell differentiation	10	4	0.13	5137	6.0e-06	6.0e-06	0.944	0.94361
GO:0007160	cell-matrix adhesion	171	11	2.27	5195	1.9e-05	3.4e-05	0.969	0.60792
GO:0043524	negative regulation of neuron apoptotic process	152	10	2.02	359	3.7e-05	3.7e-05	0.025	0.02470
GO:0007411	axon guidance	168	11	2.23	244	1.6e-05	5.1e-05	0.012	0.00135
GO:0042698	ovulation cycle	93	8	1.23	5317	3.3e-05	7.4e-05	1.000	0.54108
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	998	31	13.24	14	1.0e-05	0.00013	2.8e-06	2.2e-06
GO:0048485	sympathetic nervous system development	23	4	0.31	5139	0.00022	0.00022	0.944	0.94419
GO:0008285	negative regulation of cell proliferation	585	23	7.76	316	3.8e-06	0.00023	0.020	0.00057
GO:0035590	purinergic nucleotide receptor signaling pathway	10	3	0.13	5291	0.00026	0.00026	1.000	0.99990
GO:0030574	collagen catabolic process	24	4	0.32	5119	0.00026	0.00026	0.939	0.93882

Table 9: GO enrichment for DE genes with the same direction in both B10.D2 mouse strain and rat

Regarding LncRNAs, as expected we observed modest syntenic conservation, i.e. LncRNAs in equivalent genomic positions between species. 147 LincRNAs were identified in the same relative position of the same closest adjacent ENSEMBL gene. Out of these 4 had a pain gene as their closest genomic feature, table 10.

lincRNA	ENSEMBL_id	Distance	symbol
chr2:55315059-55415131(+)	ENSMUSG00000026824	20839	Kcnj3
chr2:75673219-75673509(-)	ENSMUSG00000015839	-2004	Nfe2l2
chr14:68043810-68081656(+)	ENSMUSG00000022055	2207	Nefl
chr14:70872255-70874020(+)	ENSMUSG00000022103	16110	Gfra2

Table 10: Syntenically conserved LincRNAs between rat and mouse with the same pain gene as their closest gene

Moreover the only syntenically conserved LincRNA with higher than random and significant correlation to its adjacent gene in rat is the LincRNA chr1:44859521-44866436(+) which is close and correlated to Oprm1. The LincRNA is upstream of the gene and is significantly downregulated as is the protein coding gene. Oprm1 is an opioid receptor like Oprd1, associated with pain intensity and opioid analgesia and it also interacts with Oprd1 forming an heterodimer. Thus we have identified two pairs of significantly DE LincRNAs, highly correlated with opioid receptors Oprm1 and Oprd1, in rat and mouse respectively. These opioid receptors in turn interact to form heterodimers. We should note that the fold change of the opioid receptor and the gene is the same, i.e. down-regulation after SNI surgery, in both species, an indication that the LincRNAs might induce gene expression.

Regarding antisense LncRNAs, as expected, we identified both Scn9a and Kcna2 antisense LncRNAs in both species and strains. Neither these antisense transcripts are part of the ENSEMBL/GENCODE or RefSeq annotation. Kcna2 antisense does not reach significance in rat nor in mouse, Scn9a does reach significance in rat but not in mouse after SNI surgery, as also reported in (Koenig et al., 2015), table 11.

LncRNA	Gene ID	Gene symbol	Lnc lfc	Lnc p,value	Gene lfc	Gene p.value	CPC score	Organism
chr2:229296258-229305952(-)	ENSRNOG00000018285	Kcna2	-0.669228825	0.2201968759	-1.118629743	2.325746E-06	-0.977888	Rattus Norvegicus
chr3:107107205-107115062(-)	NM_008417	Kcna2	-0.712913105	0.1374298266	-0.332733876	0.0020456076	0.447681	Mus musculus BALB/c
chr3:107107205-107115062(-)	NM_008417	Kcna2	-0.30151431	0.785324861	-0.25723331	0.0196038072	0.447681	Mus musculus B10.D2
chr3:59057403-59276093(+)	ENSRNOG00000006639	Scn9a	-2.357018731	1.425246E-09	-1.08666881	1.445246E-05	-0.815863	Rattus Norvegicus
chr2:66634323-66642309(+)	ENSMUSG00000075316	Scn9a	0.0687529985	0.9926893542	-0.145395638	0.2883473421	-1.12072	Mus musculus BALB/c
chr2:66634323-66642309(+)	ENSMUSG00000075316	Scn9a	-0.419457319	0.9347317631	-0.022210573	0.9319527903	-1.12072	Mus musculus B10.D2

Table 11: Conserved antisense LncRNAs on the opposite strand of Scn9a and Kcna2 genes

In general we identified 388 syntenically conserved antisense LncRNAs on the opposite strand of orthologous genes in mouse and rat. Moreover we identified 25 syntenically conserved LncRNAs antisense of pain genes, table 12. Very important pain genes were found to have an antisense non-coding RNA transcribed in both species. Moreover Fyn, Prkca, Scn9a, Trpm3 and Atp1b3 pain genes have an antisense LncRNA with opposite expression patterns.

LncRNA	Gene ENSMEBL ID	Gene symbol	cpc
chr10:39555935-39560136(-)	ENSMUSG00000019843	Fyn	-0.472081
chr11:55394494-55395613(+)	ENSMUSG00000018593	Sparc	-1.05264
chr11:63150892-63151309(-)	ENSMUSG00000018217	Pmp22	-0.886487
chr11:70239888-70246096(+)	ENSMUSG00000000320	Alox12	0.125631
chr11:81966759-81995840(+)	ENSMUSG00000020704	Asic2	-0.833933
chr11:102739325-102762503(-)	ENSMUSG00000020926	Adam11	0.539235
chr11:107935614-107937468(+)	ENSMUSG00000050965	Prkca	-0.893234
chr13:112505712-112508381(-)	ENSMUSG00000021756	Il6st	-1.01211
chr14:103778950-103851424(+)	ENSMUSG00000022122	Ednrb	0.298243
chr15:101214611-101225267(-)	ENSMUSG00000000531	Grasp	-0.116895
chr16:87934064-87936434(+)	ENSMUSG00000022935	Grik1	-1.01687
chr16:92690953-92693516(+)	ENSMUSG00000022952	Runx1	-1.09998
chr19:6969343-6970896(+)	ENSMUSG00000024960	Plcb3	-0.934247
chr19:22435556-22448608(-)	ENSMUSG00000052387	Trpm3	-0.191057
chr2:66634323-66642309(+)	ENSMUSG00000075316	Scn9a	-1.12072
chr2:75671430-75690556(+)	ENSMUSG00000015839	Nfe2l2	0.169968
chr2:127481675-127485719(+)	ENSMUSG00000079056	Kcnip3	-1.07449
chr3:60782742-61004319(-)	ENSMUSG00000027765	P2ry1	0.749763
chr3:101592328-101592581(+)	ENSMUSG00000033161	Atp1a1	-1.18998
chr4:132560740-132604797(-)	ENSMUSG00000056529	Ptafr	-0.216132
chr5:43867709-43869237(-)	ENSMUSG00000029084	Cd38	-1.12311
chr6:125241923-125242339(-)	ENSMUSG00000030337	Vamp1	-0.713977
chr7:91253594-91259556(-)	ENSMUSG00000052572	Dlg2	-1.2575
chr7:114635520-114636347(+)	ENSMUSG00000030669	Calca	-1.18392
chr9:96345647-96364371(+)	ENSMUSG00000032412	Atp1b3	-1.00742

Table 12: LncRNAs antisense of pain genes, syntenically conserved between mouse and rat

Discussion

In this study we have studied transcriptional changes in mouse DRGs after SNI surgery without being restricted by the annotated gene models found in both major genomic annotation consortia. As our aim was not to perform a complete reconstruction of the transcriptome, we only selected a sub-set of predictions on which we could have more confidence based on the expression consistency, strength, pattern and differential expression between conditions of interest. Of course all results we presented, regarding LncRNAs, have derived from predicted models and thus require Q-PCR validation in the wet lab in other samples to assess biological reproducibility.

In this study of the SNI model of peripheral neuropathy, we used mouse strains with significantly different responses and intensity of induced mechanical hypersensitivity after the pain surgery. We found that the high pain strain BALB/c, has more genes significantly DE between SNI and sham operated animals than the low pain strain B10.D2. The high pain strain had more prominent transcriptional changes in voltage gated potassium and sodium ion channels indicating that different responses on the dysregulation of genes encoding these channels play significant role in maintaining neuropathic pain and differentiating its severity between strains. BALB/c strain had a very similar profile of transcriptional changes regarding ion channels and pain genes to rat, on the other hand B10.D2 strain did not have significant responses of these genes differentiating conditions. Moreover rats that underwent the SNT pain model and had significantly induced mechanical hypersensitivity were much more similar to the high pain strain than the low pain strain.

Biological processes of axon guidance, regulation of neuronal regeneration and development as well as regulation of ion channels, signalling and response to stimuli, learning and memory are highly enriched amongst the biological process related to genes significantly DE in both mouse strains. These processes are essential for the phenotype of neuropathic pain. Regarding genes with significant different responses

between strains we found that these are highly enriched for biological processes related to axon guidance and neuron development and regeneration, as well as potassium ion transport, immune system response and chemokine regulation. Moreover genes related to neuronal plasticity and calcium channels were also enriched amongst the genes with significantly different response between the high and low pain strain.

The above findings were extended by the identification of 4034 LncRNAs, expressed in mouse DRG. These non-coding transcripts that are mostly bi-exonic are expressed 10-times lower than protein coding genes but nevertheless they have a consistent expression pattern that could separate samples according to the biological condition, namely SNI or sham surgery, BALB/c or B10.D2 strain. They were found to be significantly DE SNI vs sham in very similar percentages to those of known genes in BALB/c and B10.D2 strain. 171 DE LncRNAs were found to be significantly dysregulated in the high pain strain and 124 in the low pain strain. As expected, we have also found modest syntenic conservation between rat and mouse predicted LncRNAs

Out of the predicted antisense LncRNAs, 8 are significantly DE and antisense of significantly DE genes. Moreover two of them have an opposite expression pattern, similar to that of Kcna2 (Zhao et al., 2013) and Scn9a (Koenig et al., 2015) antisense transcripts. The pair of protein coding gene Nalcn - antisense LncRNA was only found significantly DE in the high pain strain. Nalcn, a sodium leak channel, is important for the regulation of neuronal excitability by contributing a basal Na+ leak conductance in neurons (Ren, 2011). Thus Nalcn might be regulated by this antisense LncRNA and contribute to the different levels of pain intensity after the SNI surgery between strains. The same is true also for Tshz2 gene, which is related to neuron development, and its antisense LncRNA with opposite expression pattern found to be significantly DE only in the high pain strain. Amongst antisense LncRNA we have also identified, the published but not included in the ENSEMBL or RefSeq annotation, LncRNAs antisense of

Kcna2 and Scn9a. In general 60 LncRNAs were predicted antisense of pain genes.

Regarding intergenic LincRNAs most of them were in medium to close proximity to known genes and those found to be relatively close to known genes were also enriched for significantly DE. 53 were significantly DE with a significantly DE closest genomic neighbour in BALB/c strain and 47 in B10.D2. 4 of them were highly and significantly correlated with their closest protein coding pain gene. These pairs of LincRNA and highly correlated closest pain genes include a transient receptor potential and a potassium channel, a histamine receptor and an opioid receptor. Moreover, there is a significantly DE LincRNA close to the significantly DE Oprd1 opioid receptor in mouse, where in rat there is a significantly DE LincRNA closest to, the partner of Oprd1, Oprm1 in the same upstream position of the opioid receptor and with the same direction of change – down-regulation. Thus these LincRNAs might regulate *in-cis* those opioid receptors that are significantly DE and functionally important for neuropathic pain.

In this study we have efficiently used RNA-sequencing to predict LncRNAs and at the same time to quantify transcription changes in mouse DRGs between SNI and sham operated animals, but also between strains of high and low mechanical hypersensitivity. We have summarized our results by reporting enriched biological processes related to neuropathic pain and also found that novel LncRNAs are putative mediators of neuropathic pain. We have computationally identified a subset of intergenic and antisense LncRNAs, which given their genomic context, DE and expression pattern suggests they might be functionally important for maintaining neuropathic pain or differentiating its intensity.

As LncRNAs are not highly conserved across species and are highly tissue specific we should be very careful in translating our findings to human or other species. On the other hand some very important LncRNAs have been highly tissue or developmental stage specific but conserved across species. In this study the usage of animal models of pain gave us the

ability to study the transcriptional profile and identify LncRNAs in the highly pain relevant tissue of DRG. As harvesting of DRGs under well induced pain states would have been impossible without animal models of pain we believe that the usage of animal models gave us valuable insights regarding the response to peripheral nerve injury and the molecular changes involved. Moreover we found some conserved LncRNAs between mouse and rat DRG and we will further study if these are conserved in other species.

We should note that all results of DE LncRNAs are based on predictions from RNA-seq data of a biologically relevant experiment. Future work of this project will involve functional validation of these targets as well as integration of more data types from high-throughput assays regarding chromatin modifications and transcription start-site prediction.

Clustering of patients with diabetic neuropathy reveals distinct neuropathic pain dimensions

Overview

In this chapter we present methods and results from analysing clinical data and self-reported quality of life questionnaires data from patients with diabetes mellitus suffering from painful or painless neuropathy. Our aim is to identify distinct sensory profiles or pain qualities using self-reported questionnaires and in general assess the ability of these tests to capture the phenotype of diabetic neuropathy observed in our dataset. We will also associate these sensory profiles with results from quantitative sensory testing and clinical markers and also test self-reported questionnaires for agreement between their results.

Introduction

As the International Association for the Study of Pain (IASP) defines it, Neuropathic Pain is “pain initiated or caused by a primary lesion or dysfunction in the nervous system.” (Treede et al., 2008). Thus, for the diagnosis of neuropathic pain an underlying disease or lesion on the somatosensory system, which has in turn neuropathic pain as a symptom, should be identified. Particularly, patients suffering from diabetes mellitus exhibit diabetic peripheral neuropathy pain in percentages ranging from 28-49%; and of those patients 25-50% develop a neuropathic pain phenotype (Themistocleous et al., 2016). Thus painful neuropathy is one of the most frequent complications of diabetes.

In order for the optimal treatment to be delivered, neuropathic pain must be correctly diagnosed. This can be a difficult task as this kind of pain is due to highly heterogeneous clinical conditions. For the accurate diagnosis of neuropathic pain numerous subjective pain questionnaires and standardised sensory tests have been developed. Due to this heterogeneity of causes and clinical symptoms there is an emerging need to classify patients

of neuropathic pain accordingly and provide them with the most effective treatment as soon as possible. As the developers of The Neuropathic Pain Symptom Inventory (NPSI) questionnaire state “In this context, we thought it would be of interest to develop and validate a specific self-questionnaire for the assessment of the different symptoms of neuropathic pain. Ideally, such a questionnaire could represent a useful and exploitable tool for large cohorts of patients in multicentre studies and give information comparable to that provided by quantitative evaluation, as regards the nature and intensity of the various painful symptoms. ” (Bouhassira et al., 2004) .

The Neuropathic Pain Symptom Inventory (NPSI)

The approach described above gave birth to the NPSI self-completed pain questionnaire, presented in 2004 and designed to evaluate the different symptoms of neuropathic pain and to assess its intensity. NPSI is not a screening tool for neuropathic pain but its main goal is rather to measure psychometric properties that can be effectively used to classify patients with painful neuropathy who may have differential response to treatment. The test has been proven to have good construct validity, precision and recall. The English version of NPSI is comprised of 12 questions, which are grouped together in order to give the following 5 sub-scores: Burning (superficial) spontaneous pain derived from question 1 (Q1), Pressing (deep) spontaneous pain derived from Q2 and Q3, Paroxysmal pain derived from Q5 and Q6, Evoked pain derived from Q8, Q9 and Q10, Paresthesia/dysesthesia derived from Q11 and Q12 (Bouhassira et al., 2004). The sum of these sub-scores gives the total score and the average the average total score. For each question the patient has to respond in a numeric quantitative scale from 0 to 10, where 0 means no pain and 10 means the worst pain imaginable. There are also two questions which can be answered in an ordinal categorical scale, namely “During the past 24 h, your spontaneous pain has been present for how many hours?” and “During the past 24 h, how many of these pain attacks have you had?” which aim to evaluate spontaneous paroxysmal and ongoing pain. Given these 5 sub-scores, NPSI has been found to be able to distinguish between five distinct

clinical dimensions of neuropathic pain which are also relevant to response to treatment.

An example of a digital version of NPSI can be seen in figure 1. Although recent studies have used data analysis techniques in order to identify clusters of patients with distinct pain signatures, or in other words dimensions of neuropathic pain or sensory profiles (Freeman et al., 2014), there are different pain profiles emerging under different conditions causing neuropathic pain.

Figure 1: A digital version of the NPSI questionnaire as it is in the database used by the current study.

Douleur Neuropahique en 4 Questions (DN4)

A year after introducing NPSI, the French Neuropathic Pain Group introduced a new clinician-administered questionnaire comprising of only 4 questions, thus the name DN4, which has been found able to distinguish between patients suffering from neuropathic pain versus non-neuropathic pain (Bouhassira et al., 2005). Thus DN4 is a screening tool for neuropathic pain. This small and simple questionnaire, comprising of a small number of features sufficient to identify neuropathic pain, has great sensitivity and specificity in distinguishing between non-neurological lesions and neurological lesions. DN4 is a clinician-administered questionnaire. It is deliberately very small and simple, so it can be used both by specialists and non-specialists, on a very large scale, locally or remotely. DN4 and NPSI are complementary tools. In terms of the scoring method, DN4 has 4 groups of questions where only a binary response is allowed. A positive response scores 1 and a negative scores 0. If the sum of all questions in these 4 groups, ranging from 0 to 10, is above 4 then DN4 classifies the patient as suffering from neuropathic pain. An English version of the DN4 questionnaire, as used in our study, is in figure 2.

Diagnosing Neuropathic Pain - DN4 Questionnaire

Please complete this questionnaire by ticking one answer for each item in the 4 questions below.

Interview Of The Patient

Question 1: Does the pain have one or more of the following characteristics?

1: Burning	○ Yes	○ No
2: Painful Cold	○ Yes	○ No
3: Electric Shocks	○ Yes	○ No

Question 2: Is the pain associated with one or more of the following symptoms in the same area?

4: Tingling	○ Yes	○ No
5: Pins and Needles	○ Yes	○ No
6: Numbness	○ Yes	○ No
7: Itching	○ Yes	○ No

Examination Of The Patient

Question 3: Is the pain located in an area where the physical examination may reveal one or more of the following characteristics?

8: Hypoesthesia to touch	○ Yes	○ No
9: Hypoesthesia to prick	○ Yes	○ No

Question 4: In the painful area, can the pain be caused or increased by:

10: Brushing ?	○ Yes	○ No

TO COLLATE:

- Score 1 to each **YES** answer
- Score 0 to each **NO** answer
- If the score is 4 or higher then the pain is **likely** to be **neuropathic** pain.
- If the score is less than 4 then the pain is **unlikely** to be neuropathic pain.

Figure 2: The English form of the DN4 questionnaire.

Toronto clinical scoring system (TCSS)

The TCSS is a simple screening tool, similar to DN4, that has been developed in order to asses the presence and severity of diabetic poly neuropathy (Bril and Perkins, 2002). TCSS is completed after a simple neurological examination on the foot and the upper limb. More specifically it assesses the presence of symptoms of pain, numbness, tingling and weakness on the foot and also ataxia and upper limb symptoms. Then a score of 1 is assigned for the presence of each syndrome. In addition, it assesses the response of tendon reflexes on both left and right side using a scale of 0 for normal, 1 for reduced and 2 for absent reflexes. Finally a series of sensory tests including pinprick, temperature, light touch, vibration and position are carried out and a score of 0 is assigned for normal sensation and 1 for abnormal. Again these tests are carried out for the right and left side. An extra binary score (0-1) indicates if the patient has underlying neuropathy. Then the following sub-scores are calculated: Symptoms, Reflex, Sensation. Those scores can be adjusted, when the indication for neuropathy has been taken into account, or un-adjusted when not. The total score is the sum of all scores and ranges from 0 to 19. A TCSS questionnaire as used in our study is in figure 3.

Foot Symptoms

Pain

○ Absent ◉ Present ○ unknown

Numbness

○ Absent ◉ Present ○ unknown

Tingling

○ Absent ◉ Present ○ unknown

Weakness

◉ Absent ○ Present ○ unknown

Other Symptoms

Ataxia

◉ Absent ○ Present ○ unknown

Upper-limb symptoms

◉ Absent ○ Present ○ unknown

	Right (TCSS Normal=0 -Reduced=1 Absent=2)				Left (TCSS Normal=0 -Reduced=1 Absent=2)			
Patella	◉ normal	○ reduced	○ absent	○ ND	◉ normal	○ reduced	○ absent	○ ND
Achilles	○ normal	○ reduced	◉ absent	○ ND	○ normal	○ reduced	◉ absent	○ ND

Sensation	Right			Left		
Pinprick	○ Normal	◉ Abnormal	○ ND	○ Normal	◉ Abnormal	○ ND
Temperature	◉ Normal	○ Abnormal	○ ND	◉ Normal	○ Abnormal	○ ND
Light Touch	◉ Normal	○ Abnormal	○ ND	◉ Normal	○ Abnormal	○ ND
Vibration	◉ Normal	○ Abnormal	○ ND	◉ Normal	○ Abnormal	○ ND
Position	◉ Normal	○ Abnormal	○ ND	◉ Normal	○ Abnormal	○ ND

Figure 3: The TCSS pain questionnaire

The Quantitative Sensory Testing (QST)

QST is a standardised protocol of sensory testing which assesses evoked perception in response to a standardised stimulus developed by the German Neuropathic Pain Network (DFNS) in 2006 (Rolke et al., 2006) . QST protocol involves assessment of heat and mechanical detection and pain thresholds. More specifically mechanical detection threshold is assessed by Von Frey filaments and a 64 Hz tuning fork, pain threshold is assessed by pinprick stimuli and blunt pressure and heat detection thresholds with computerized generators of thermal stimuli. Sub-scores for Cold and Heat Detection Thresholds (CDT, HDT), Mechanical Detection Threshold (MDT), Mechanical Pain Threshold (MPT), Mechanical Pain Sensitivity and Pain Allodynia (MPS, ALL), Wind-Up Ratio (WUR), Vibration Detection Threshold (VDT) and Pressure Pain Threshold (PPT) are calculated and z-transformed. Given these 13 sub-scores, 26 distinct hyper- or hypo-phenomena can be measured. Moreover QST results can be used for group comparisons and the identification of somatosensory phenotypes. The DFNS consortium has provided expected distributions of measurements, standardized procedures and reference values.

The 7-Day pain diary

The 7-Day Pain Diary is a record keeping log where patients record the intensity and time of pain they experience throughout the day. Pain diary records pain for 7 days and patients are asked to to complete it from 9am to 9pm throughout the day. The scoring scale is 0 for “no pain” to 10 for “the worst pain imaginable”. Mean average scores for the 7-day period and the standard deviation are usually calculated. Patients are also asked to shade in body-maps where they experience pain as well as any pain treatment medication they might take. Pain diaries can be completed online, using specialised mobile apps or administered in hard-copy form.

Clinical markers

We also identified correlations between scores and groups of patients classified according to quality of life questionnaires and clinical variables. More specifically we have examined relationships between different questionnaire derived scores and HbA1c blood test, Intra Epidermal Nerve Fibre Density (IENFD), the age, gender and Body Mass Index (BMI) of patient, the duration of diabetes and neurological examinations such as the MRC sensory test, the Warm and the Cold Sensibility Index (WSI – CSI).

HbA1c is one of the main blood tests for diabetes. HbA1c measures glycated haemoglobin and provides an amassment of long-term glycemic control as it gives an indication of average glucose levels for a period of two to three months. HbA1c has been found to be correlated with the risk of long-term diabetes complications (Khan et al., 2016). HbA1c is measured in mmol/mol or in percentage and the normal levels are below 42mmol/mol or below 6% respectively.

IENFD is a neurological examination which identifies the density of small nerve fibres. IENFD provides an assessment of small-fibre neuropathy, like diabetic neuropathy. The consensus method involves skin biopsy using a 3mm circular punch tool, usually at the ankle, and then the nerve density is quantified in fibres per mm. A low IENFD indicates small-fibre neuropathy and is associated with a higher probability of developing neuropathic pain but it cannot be correlated with the intensity of it (Lauria et al., 2010).

The MRC sensory score is a standardised sensory score derived from neurological examination as described by the Medical Research Council. The WSI and CSI scores are defined as: WSI = (warm pain detection threshold − warm threshold)/(warm pain detection threshold − reference temperature) and CSI (cold pain detection threshold − cold threshold)/(cold pain detection threshold − reference temperature) (Themistocleous et al., 2016).

Methods

The main goal of this study was to identify distinct dimensions/qualities of neuropathic pain which can optimally separate patients according to their respective sensory profiles, pain symptoms and severity. We aimed to find pain qualities which correlated well with pain intensity.

Imputing missing values

Most of the algorithms used and especially PCA have difficulty handling missing values, so we either need to discard incomplete observations, i.e. patients or samples with any missing value, or impute missing values. In studies where we have plenty of data points per sample instead of a big number of individual samples, an efficient strategy for imputing missing values is necessary. We can always use the simple strategy of imputing missing values with the mean value for the specific variable, but usually this is not advised. Although this may preserve some descriptive statistics like the mean and the standard deviation it can also seriously affect the downstream analysis and inference statistics (Josse and Husson, 2012). Instead we have used a method for handling missing values with multivariate data analysis. In this way missing values were imputed by using a Principal Component Analysis (PCA) model. The number of principal components used was estimated by cross-validation, each cell/datapoint, or a certain percentage of cells, of the data matrix is alternatively removed and predicted with a PCA model using different dimensions ranging from a minimum to a maximum number. The number of components which led to the smallest mean square error of prediction (MSEP) was retained. Next, missing values were completed first by the mean of the variable and then by iterative PCA steps until convergence of the algorithm when the PCs were not changed by the imputation of the missing value. The R package missMDA (Josse and Husson, 2012) has been used for this process.

Clustering

To optimally partition data in a way that more similar objects were grouped together we performed clustering. In this study we used various forms of clustering but mostly unsupervised k-means clustering. We have usually clustered data not according to the original values but according to some transformed values by a certain function, like log2, PCA or the varimax rotated loadings. A brief overview of the k-means clustering algorithm is as follows:

K-means clustering (Hartigan and Wong, 1979) selects K centroids (K rows of the data matrix chosen at random) and assigns each data point to its closest centroid, then by iterative steps it recalculates the centroids using the average of all data points in a respective cluster and consequently assigns data points to their closest centroids. Iterations stop and the algorithm ends when all observations cannot be further reassigned or the maximum number of iterations has been reached.

As k-means clustering depends on the predefined number of clusters we used the following method for defining the optimal number of clusters: We plotted the within clusters sum of squares, i.e. the sum of the squared differences of each observation of the group from the group mean against the number of clusters. Where we observed a distinct drop, i.e. elbow, in the within groups sum of squares when we moved from a solution with a certain number of clusters to another we could identify the best fit.

Data Analysis and statistical tests

We have used exploratory data analysis techniques and unsupervised clustering in order to identify the optimal number of clusters which separate individuals and at the same time reduce the dimensionality of the problem. We have clustered patients and shrunk our data using data points related with the NPSI pain questionnaire and QST. Then we inspected how the loadings of a sub-set of principal components calculated from the above data analysis could be associated with pain intensity, sensory profiles and clinical variables. We have also assessed how these clusters of patients

correlate with scores from other neuropathic pain tools and clinical variables.

First our analysis involved converting raw data points in a form that could be effectively analysed. This data wrangling process involved recoding of categorical variables, scaling and normalization of quantitative ones and imputing missing values.

QST scores were first scaled and z-transformed. PCA is very sensitive to the scaling of data and to the distribution of variance (*Principal Component Analysis*, 2002) thus we carefully normalised and dealt with missing values as described above before proceeding into calculation of principal components.

For all downstream analysis we have created two different datasets from the original data: all patients and only patients with painful neuropathy. Moreover in the exploratory factor analysis of the NPSI data we divided the subset of patients with painful neuropathy into 2 datasets, patients with painful neuropathy with all NPSI data and with the two categorical variables measuring paroxysmal pain excluded.

Regarding statistical analysis we calculated Pearson's correlation for continuous numerical variables, Spearman's correlation when dealing with discrete, ordinal ranked scores and Kruskal-Walis non-parametric test to test for dependences between variables and factors. Kruskal-Walis test, essentially a non-parametric alternative to one-way ANOVA, is used in order to asses how a numeric variable or ranked score can be associated with a categorical factor with 3 or more levels. The null hypothesis of the test is that the mean ranks of the groups which are coded by the different levels of the respective factor are equal. We will reject the null hypothesis if p.values are less than 0.05. As it is a non-parametric test it does not rely on an assumption of normality. Thus it can be used where the assumptions for one-way ANOVA are not met. Particularly in our data, as published in (Themistocleous et al., 2016), only the QST z-scores were normally distributed and thus this is the only case we used one-way ANOVA to assess

the effect of numeric variables in categorical factors. In order to perform comparison between groups we used the two-tailed student's t-test.

As this is not a hypothesis driven study, but rather an exploratory analysis, we only tested for significance individual comparisons. We did not tried to prove any hypothesis by performing multiple tests, thus we did not increased the probability of type I errors. As we did not tried to answer a question having the form: what are the correlated pairs of variables under the assumption that most pairs are not correlated; we did not raised the probability of erroneously rejecting the null hypothesis just by performing multiple tests. Thus we only controlled for the comparison wise error rate (CER) and an adjustment for multiple tests was unnecessary (Bender and Lange, 2001). Results of the statistical tests presented in this chapter of exploratory analysis should only be considered as descriptive and not inferential.

Recoding variables

Recoding categorical variables was an important step of the analysis. The goal was to find a way to transform categorical variables in a way that preserves the natural order and properties of the attribute measured. In the current study we have recoded two categorical variables in the NPSI questionnaire that assesses the duration and frequency of spontaneous pain: "During the past 24 h, your spontaneous pain has been present?" with possible answers/values of "Permanently", "Between 8 and 12h", "Between 4 and 7h", "Between 1 and 3h" and "Not Available" and "During the past 24 h, how many of these pain attacks have you had?" "More than 20", "Between 11 and 20", "Between 6 and 10", "Between 1 and 5", "Not Available". In the database, these variables have the respective names spontaneous ongoing pain (NPSI_SPONTONGOING) and spontaneous paroxysmal pain (NPSI_SPONTPAROXYSMAL). As there is no obvious reason for any bias in the distribution of answers we can assume that they are uniformly distributed. Thus, we can represent hour intervals and intervals of certain counts of paroxysmal pain attacks by their mean value.

This recoding gave us the following mapping between categorical and numeric values:

Spontaneous Ongoing pain	Ordinal Value	Numerical recoding
	Permanently	24
	Between 8and 12h	10
	Between 4 and 7h	5.5
	Between 1 and 3h	2

Table 1: Recoding of Spontaneous Ongoing pain variable

Spontaneous Paroxysmal pain	Ordinal Value	Numerical recoding
	20 or more	20
	Between 11 and 20	15.5
	Between 6 and 10	8
	Between 1 and 5	3

Table 2: Recoding of Spontaneous Paroxysmal Pain variable

Transform scores into categorical variables

Subsequently we transformed the different NPSI pain sub-scores: deep spontaneous pain, evoked pain, paresthesia/dysesthesia, paroxysmal pain, superficial spontaneous pain and average NPSI total score into distinct ordinal variables of pain severity. For each of the five scores: Burning (superficial) spontaneous pain, Pressing (deep) spontaneous pain, Paroxysmal pain, Evoked pain, Paresthesia/dysesthesia, we have used the following mapping which is found in studies of various types of neuropathic pain

(Alschuler et al., 2012; Freeman et al., 2014; Miró et al., 2016) to better reflect the categorisation of pain intensity:

NPSI scores	Pain Intensity Categories
0	No neuropathic pain
(0-4)	Mild
[4-7)	Moderate
[7-10]	Severe

Table 3: Pain severity categories derived from NPSI scores

As the NPSI average total score is the average of all these five sub-scores, we assigned the same pain intensity categories to the NPSI average total score variable. We did not transform the NPSI total score, i.e. the sum of all these sub-scores.

Normalization and imputation of missing values

Bringing all variables into the same scale in order to be directly comparable is essential for downstream data analysis, as PCA is very sensitive to the distribution of variance in the dataset. We normalized and scaled all values derived from the NPSI questionnaire in a way that centred the values, mean = 0, for all variables and scaled, sd = 1, for all variables. Moreover all QST parameters have also been transformed into z-scores (mean = 0, sd=1) thus all numbers represent distance from the mean measured in the units of the standard deviation. We carried out this scaling of the data as the first step of a principal component analysis.

As we clustered data by the NPSI questionnaire we discarded all individuals which had no NPSI data, i.e. where all values of the questionnaire were not available. Thus we discarded 11 patients, 7 with painful and 4 with painless neuropathy. Any other missing values were imputed by a Principal Component Analysis (PCA) model (see Introduction/Imputing missing values).

Dataset

In order to carry out the current study and gain valuable insights regarding neuropathic pain we have used a comprehensive curated dataset of patients with diabetic neuropathy compiled by Andreas Themistocleous and

Juan Ramirez, David Bennet's group, NDCN, Oxford. The database has been maintained by Jon Lees, Orengo Group, SMB, UCL. This database consists of 191 patients' data. 80 have painless neuropathy and 111 have painful neuropathy of varying severity.

For each individual patient the database holds 945 data points, including subjective physicians diagnosis, clinical markers, data from all the standard pain questionnaires, neuropathic pain screening tools and information regarding the patient's clinical history. More precisely, the patients' data is arranged into 4 main categories, Basic Information, Physical Examination, Special Investigations, Group Definitions. Basic Information includes Patient's History, Family History, Current Symptoms and Pain Questionnaires. We will mainly focus on data regarding screening tools, pain questionnaires and standardised tests in order to identify distinct sensory profiles – qualities of pain, and to associate them with pain severity and clinical findings. Physical examination includes data from neurological examinations and group definition define whether the patient has painful or painless neuropathy.

All the data has been downloaded in the form of comma separated values flat text files (.csv) and analysed in the statistical programming environment R. We have used functions from packages "pvclust", "cluster", "dendextend", "fpc" (Galili, 2015; Hennig, 2015; Maechler et al., 2015; Suzuki and Shimodaira, 2015) to perform clustering (centroid based – K means and connectivity based - Hierarchical) and visualize results, "FactoMineR" and "missDMA" (Husson et al., 2016; Husson and Josse, 2015) to perform PCA, normalise and impute missing values and "Hmisc" (Jr et al., 2016) to calculate p.values and confidence intervals for some of the statistical tests used.

Results

Distribution of pain scores across sexes and clinical markers for patients with painful neuropathy

In the current study we analysed data from 129 males and 62 females. Out of those 76 males and 35 females had painful neuropathy according to the physicians examination. After removing the 7 patients with no NPSI data available we ende up with 72 males and 32 females. Out of these having painful neuropathy, more females suffered from severe and moderate pain relatively to males, table 4. One male patient with neuropathic pain did not score any NPSI subscore above 0, but nevertheless he is included in the painful neuropathy group as he was diagnosed with painful neuropathy according to the IASP/neuPSIG grading system (Treede et al., 2008).

NPSI scores	Gender	
	Male	Female
No - 0	1	0
Mild - (0-4)	49	15
Moderate - [4-7)	18	10
Severe - [7-10]	4	7

Table 4: Distribution of NPSI scores across genders. Only patients suffering from painful neuropathy

Although females report more severe neuropathic pain than men, Figure 4, there is no association between higher HbA1c concentration or less IENFD and the patients' gender, figure 5. Interestingly males have lower median IENFD and higher interquartile range than females. All boxplots found in this study represent the median by a thick black line, the height of the box represents the interquantile range (IQR), i.e. data points between the 1st and the 3rd quantile. Whiskers represent the extent of 3rd

quartile + 1.5 IQR and 1st quartile – 1.5 IQR. Circular dots represent outlying values which are more than 1.5 times the 3rd quartile or lower than 1.5 times the 3rd quartile.

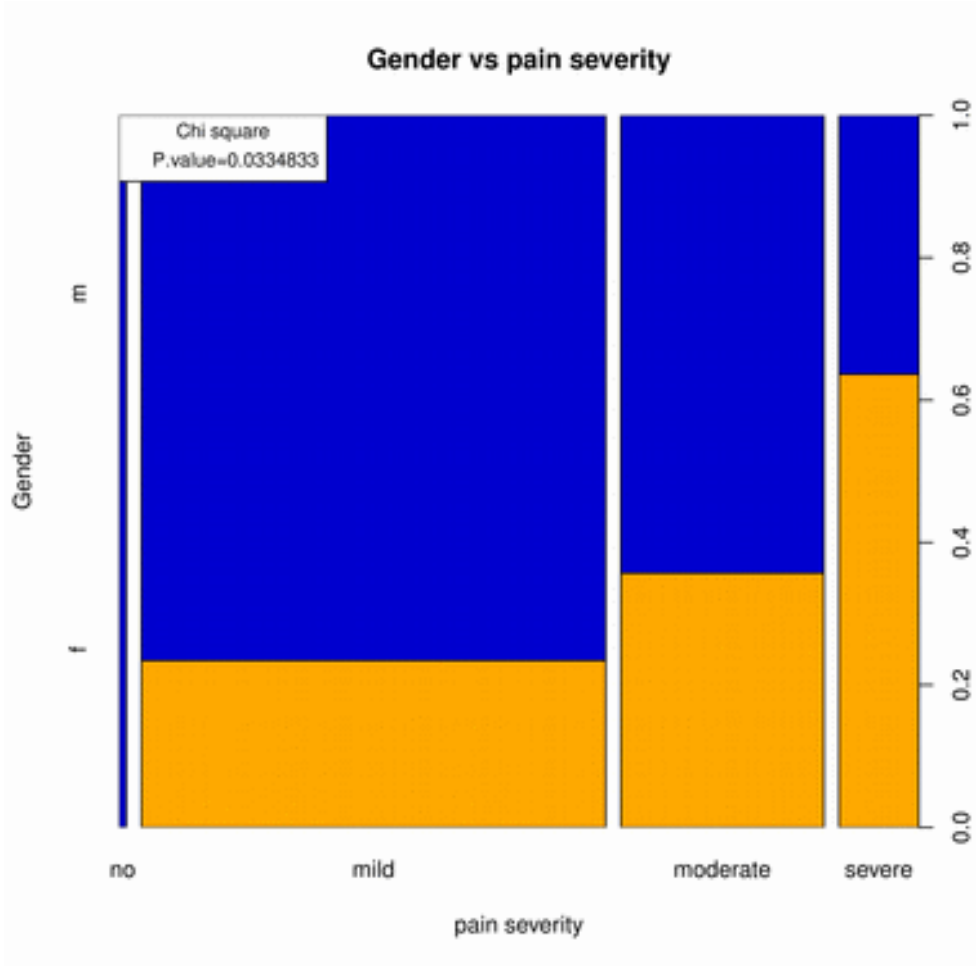


Figure 4: Females (mustard) report more severe neuropathic pain than males (blue)

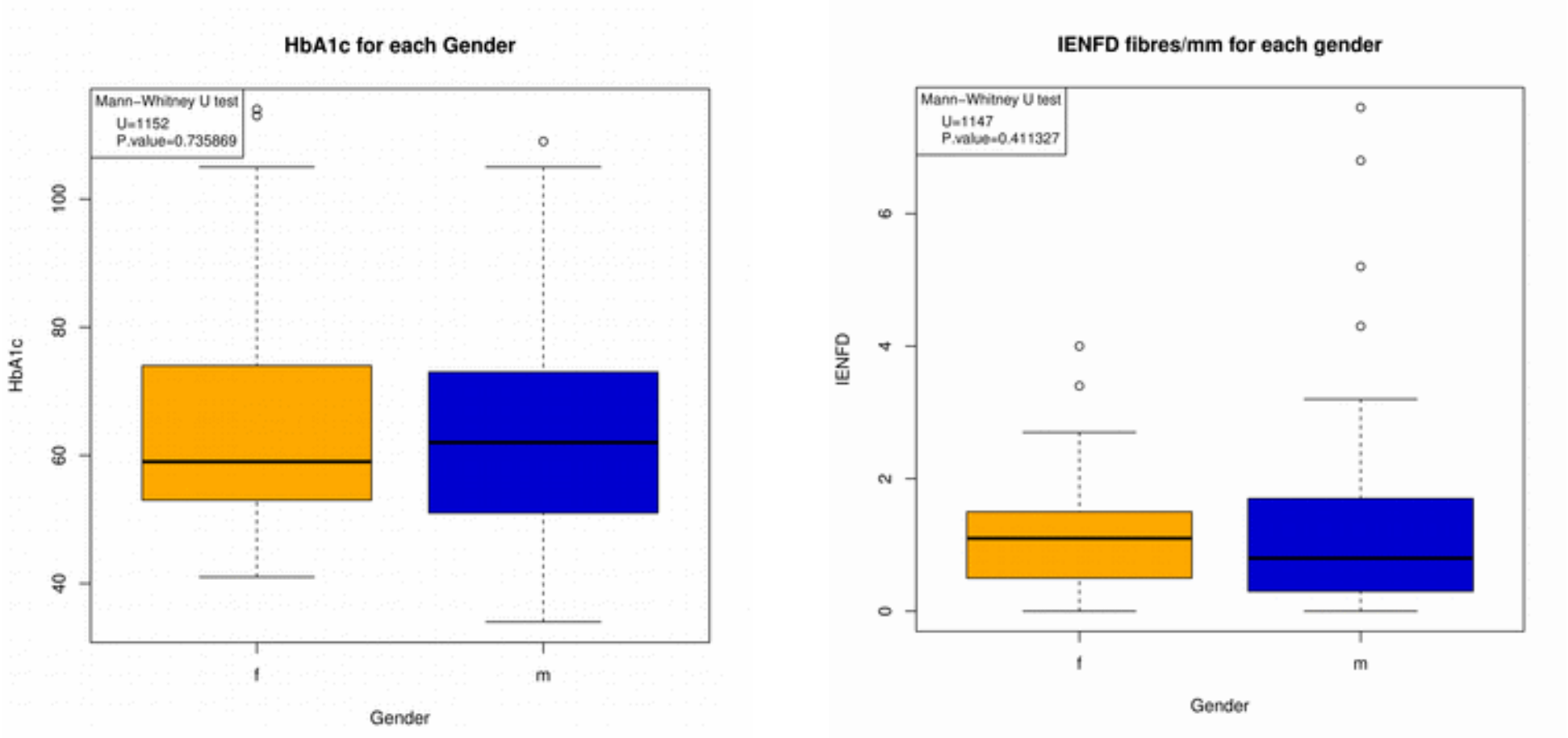


Figure 5: No association between gender and HbA1c (left) and IENFD (right).

Moreover, when we examined how NPSI total average score is correlated to clinical parameters we found significant moderate correlation between the HbA1c blood test and NPSI scores and significant mild negative correlation between the age and NPSI pain scores figure 6, but no correlation between IENFD fibres/mm and NPSI scores. Similarly there was no correlation between BMI and NPSI scores (Pearson correlation coefficient = 0.0439, p.value=0.659) and duration of diabetes and NPSI scores (Pearson correlation coefficient = −0.057, p.value = 0.566) regarding patients with painful neuropathy.

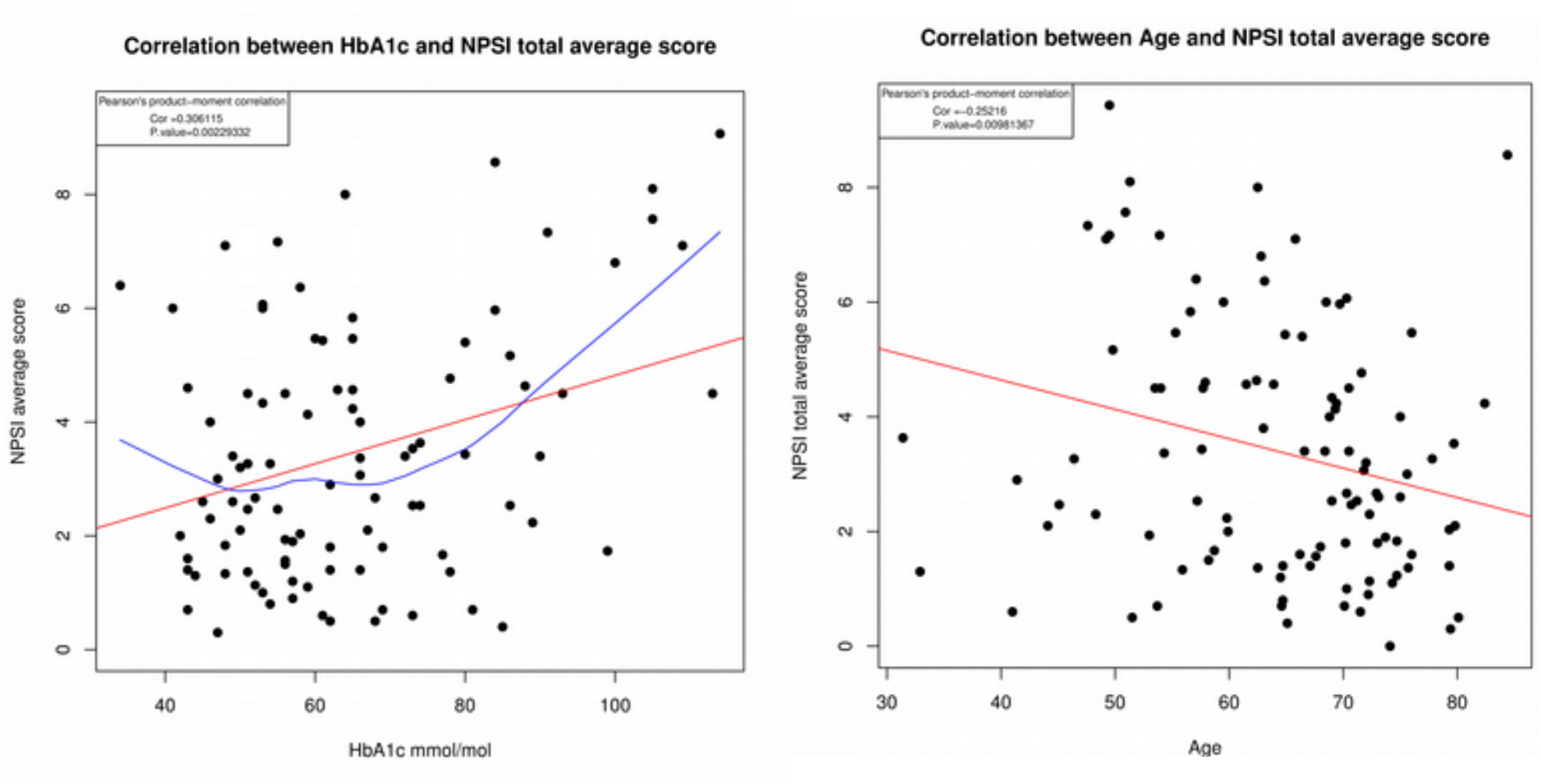


Figure 6: Significant correlations between NPSI scores and HbA1c (top) and Age (bottom). The red line is the fitted linear model line and the blue line in the HbA1c plot is the smoothed lowess curve.

Consistency of neuropathic pain screening tests

Moreover when we considered all patients, we observed very high correlations between the screening tools for neuropathic pain DN4 and TCSS, figure 7; and also between the MRC sensory score and the TCSS sensation subscore, figure 8. These results indicate that these screening tools are very efficient in classifying individuals according to the presence of neuropathic pain. This result reinforces the findings in (Themistocleous et al., 2016) where DN4 was found to have 88.3% sensitivity and 91.7% specificity. Patients scoring higher than 4 in DN4, i.e. the DN4 score threshold for neuropathic pain, have consistently higher TCSS total scores (correlation coefficient = 0.617, p.value = < 2.2e-16). This is mainly driven by the TCSS symptoms sub-score (correlation coefficient = 0.799, p.value = < 2.2e-16) and also from the TCSS sensation subscore (correlation coefficient = 0.478, p.value = 7.111e-12). Regarding DN4 score vs the TCSS reflex subscore we observed that patients with painful neuropathy show increased variance with higher IQR but median values were not significantly different.

As DN4 is mainly a screening test for the presence of neuropathic pain and TCSS is also a test for identifying the presence and severity of neuropathic pain, we hypothesized that their scoring systems might also reflect the severity of neuropathic pain in patients with painful neuropathy. Consistent with the DN4's excellent performance in the cohort of all patients, the questionnaire's total score has a strong and significant correlation (Spearman rank correlation Rho = 0.46, p.value = 7.792e−07) with the NPSI score in patients with painful neuropathy, figure 9. The same is true for the TCSS symptoms subscore (Correlation coefficient = 0.316, p.value = 0.001) but not for other TCSS scores including the TCSS total score, figure 10. This highlights that the task of classifying patients according to the presence of neuropathic pain is different to assessing the severity of it. Thus for patients with painful neuropathy only the TCSS symptom sub-score reflects pain severity according to the NPSI and DN4 scores. Moreover standardised sensory scores CSI and WSI did not show

any strong correlation with NPSI scores, except for the MRC sensory score which has lower median values for patients with severe neuropathic pain, figure 11.

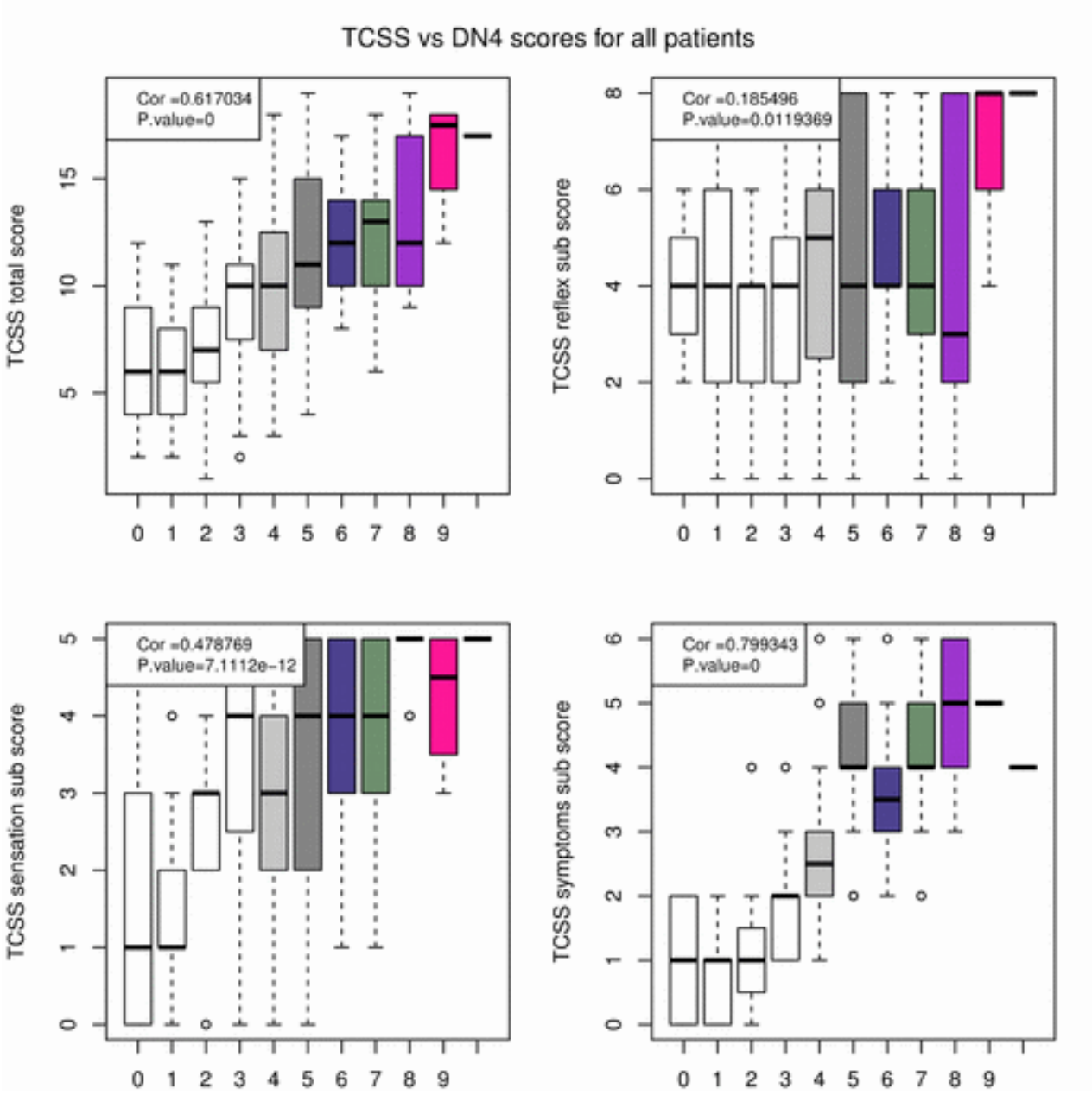


Figure 7: Correlations between TCSS scores and DN4 score. Thick black line represents the median. Box length is proportional to variance. TCSS total scores vs DN4: correlation coefficient = 0.617034, p.value = < 2.2e-16). TCSS reflex subscore vs DN4: correlation coefficient = 0.1854959, p.value = 0.01194. TCSS symptoms sub- score vs DN4: correlation coefficient = 0.7993434, p.value = < 2.2e-16. TCSS sensation subscore vs DN4: correlation coefficient = 0.4787692, p.value = 7.111e-12.

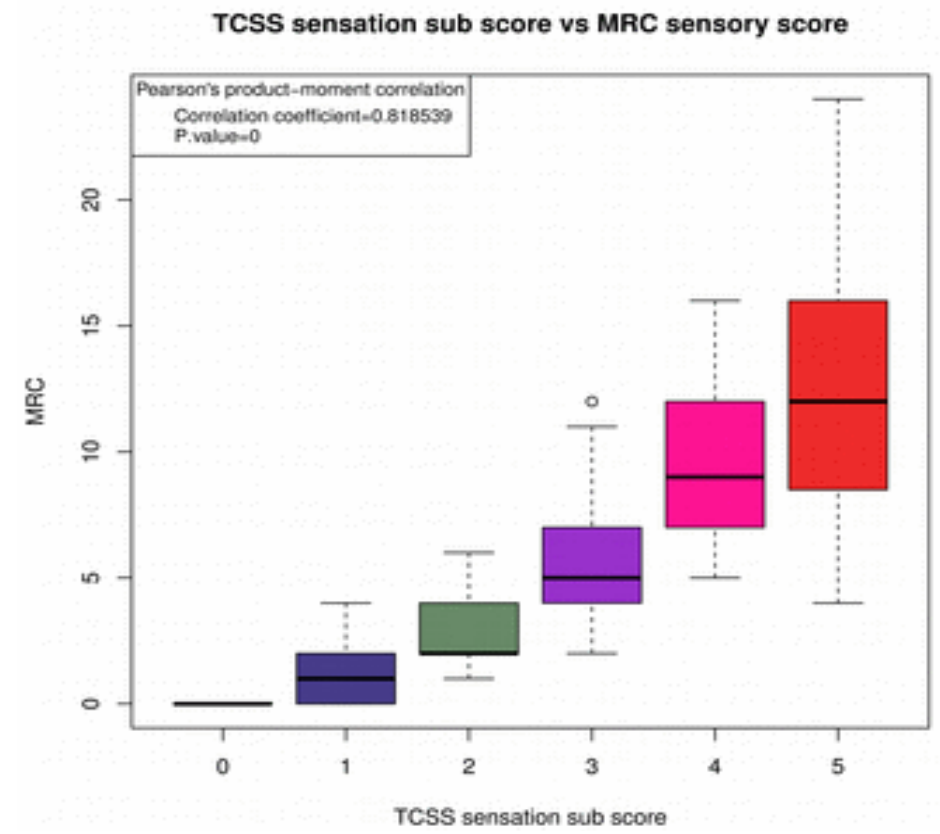


Figure 8: TCSS sensation sub score is very highly correlated to the MRC sensory score

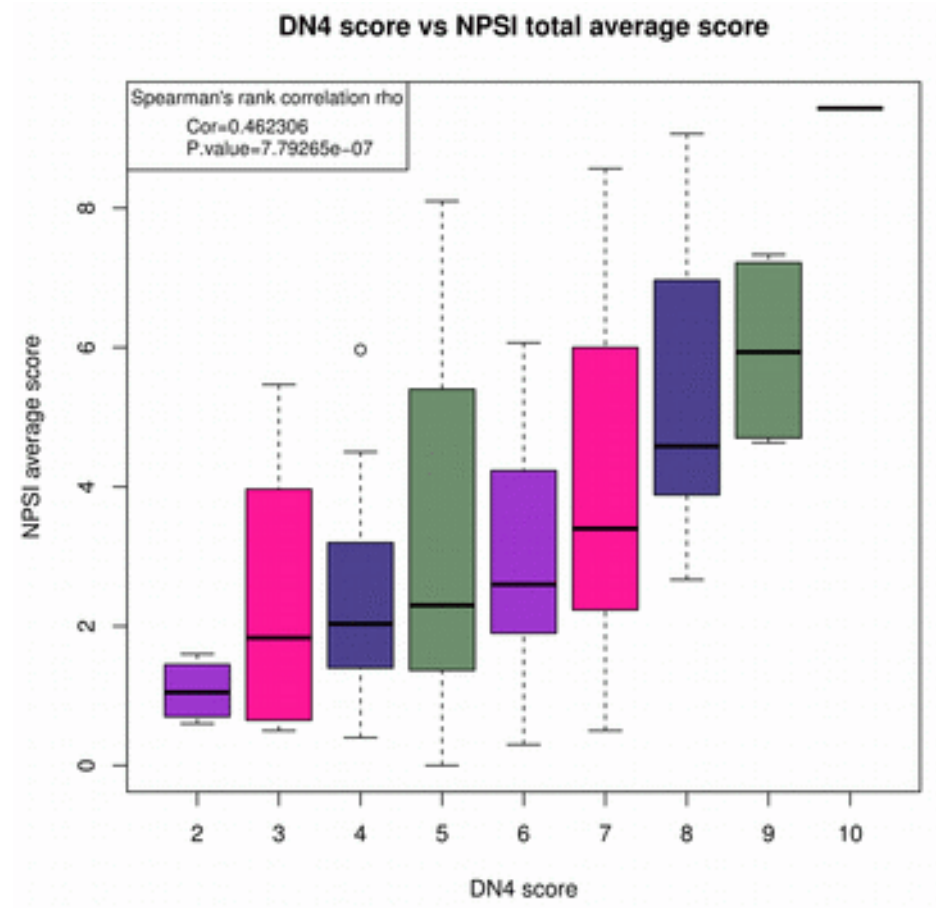


Figure 9: DN4 score is strongly correlated to NPSI average score in patients with painful neuropathy

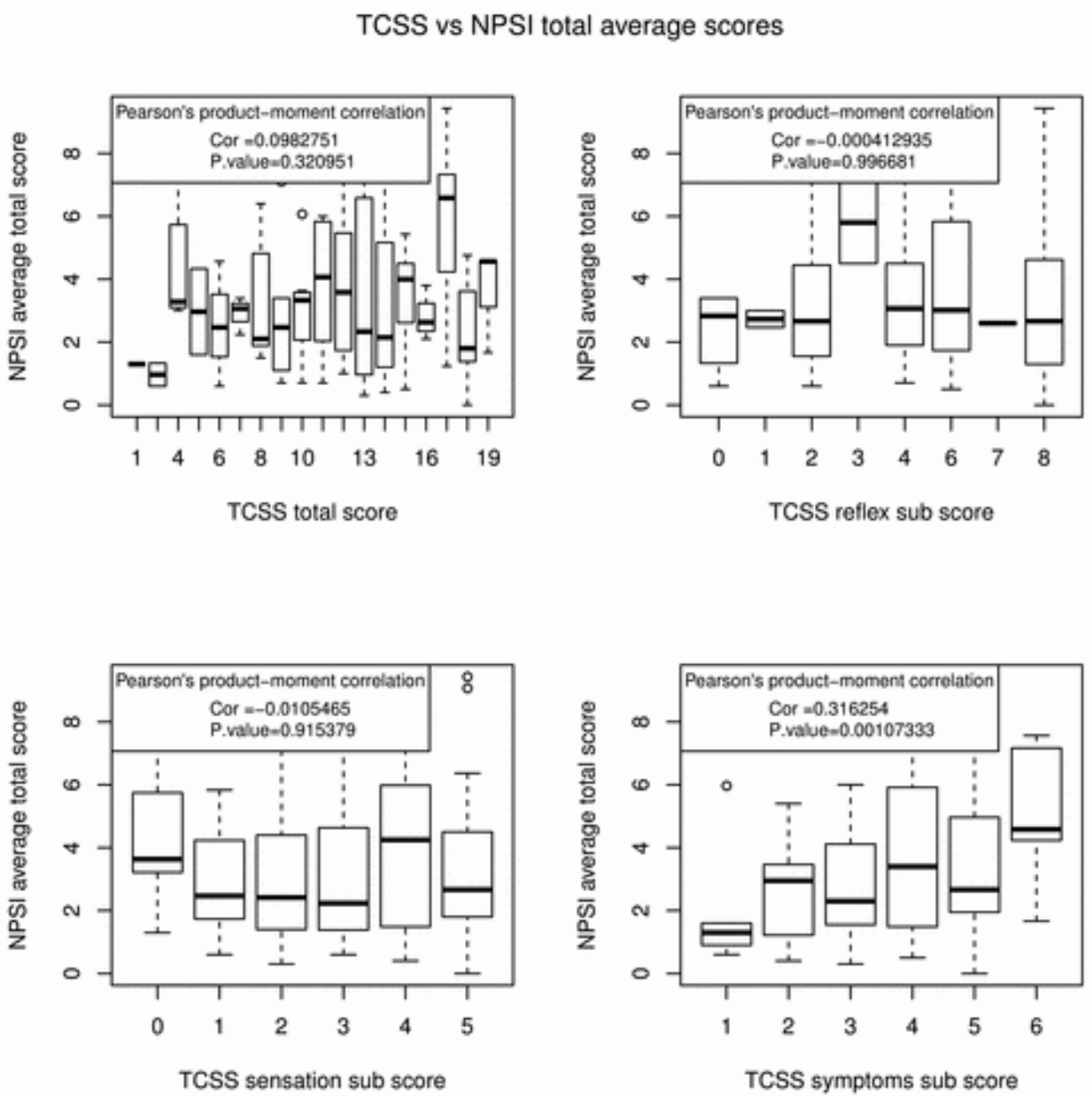


Figure 10: Correlation of TCSS scores to the NPSI average total score for patients with painful neuropathy. Only the symptoms subscore has a significant moderate correlation to the pain severity as assessed by the NPSI questionnaire.

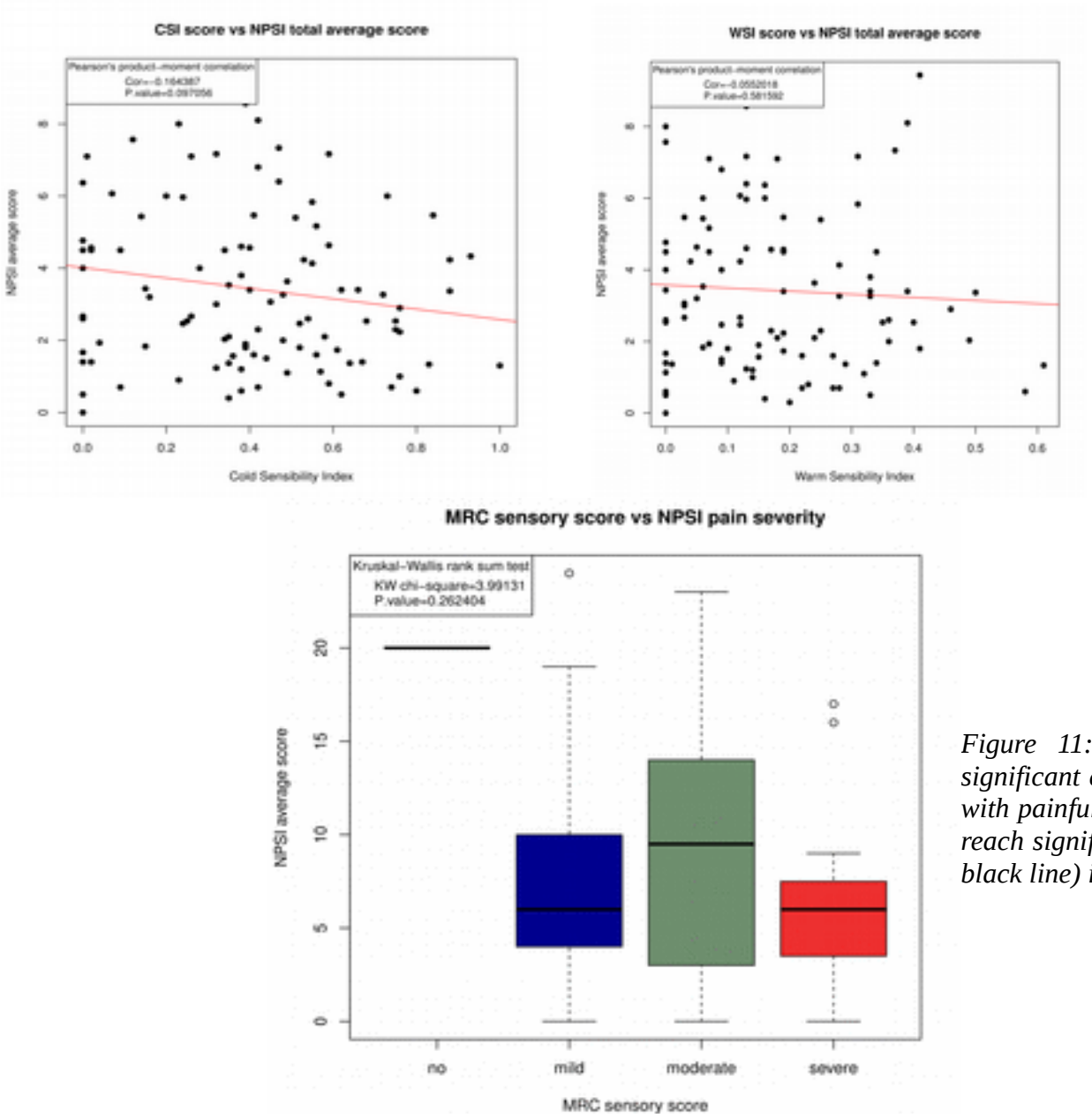


Figure 11: The WSI and CSI did not show any significant correlation with the NPSI scores for patients with painful neuropathy. The MRC sensory test does not reach significance but it has lower median values (thick black line) in patients with severe neuropathic pain.

Thus we have evidence that the NPSI average score assessing pain intensity in patients with painful neuropathy is higher in women and correlates well with the concentration of HbA1c / glycated haemoglobin. Also age has a negative correlation with pain severity. IENFD and BMI did not seem to be associated with pain severity. Moreover, although TCSS was able to classify patients according to to the presence of painful neuropathy consistently with DN4 , only DN4 is strongly associated with the severity of the neuropathic pain phenotype. On the other hand TCSS symptom subscore is moderately correlated with neuropathic pain severity. Finally MRC sensory score, highly correlated in the cohort of all patients to the TCSS sensory subscore, could not differentiate patients according to pain severity.

Quantitative Sensory Testing scores associated with self reported scores and clinical markers

All patients

Considering the complete compendium of patients with diabetes mellitus and since DN4 score was proven to have excellent sensitivity and specificity (Themistocleous et al., 2016) and was highly correlated with the TCSS results, we examined how QST was associated with the DN4 score. We observed significant and strong correlation of the DN4 score to the Mechanical Detection Threshold (MDT) and the Cold Detection Threshold (CDT) and weak correlation to the Thermal Sensory Linen and the Heat Pain Threshold, figure 12. Interestingly though, there is no strong and significant correlation of any parameter of the QST to the HbA1c which we have previously found to be strongly correlated with the severity of the neuropathic pain phenotype, figure 13. On the other hand, there is strong and significant correlation of Intra-Epidermal Nerve Fibre Density (IENFD) to the QST Mechanical Pain Sensitivity (MPS) and Mechanical Pain Threshold (MPT); moderate to Cold Detection Threshold (CPT) and Vibration Detection Threshold, figure 14. This is reasonable as better innervation leads to higher thresholds in all QST parameters but at the same

time we did not find significant correlation between reduced IENFD and increased pain intensity in patients with painful neuropathy.

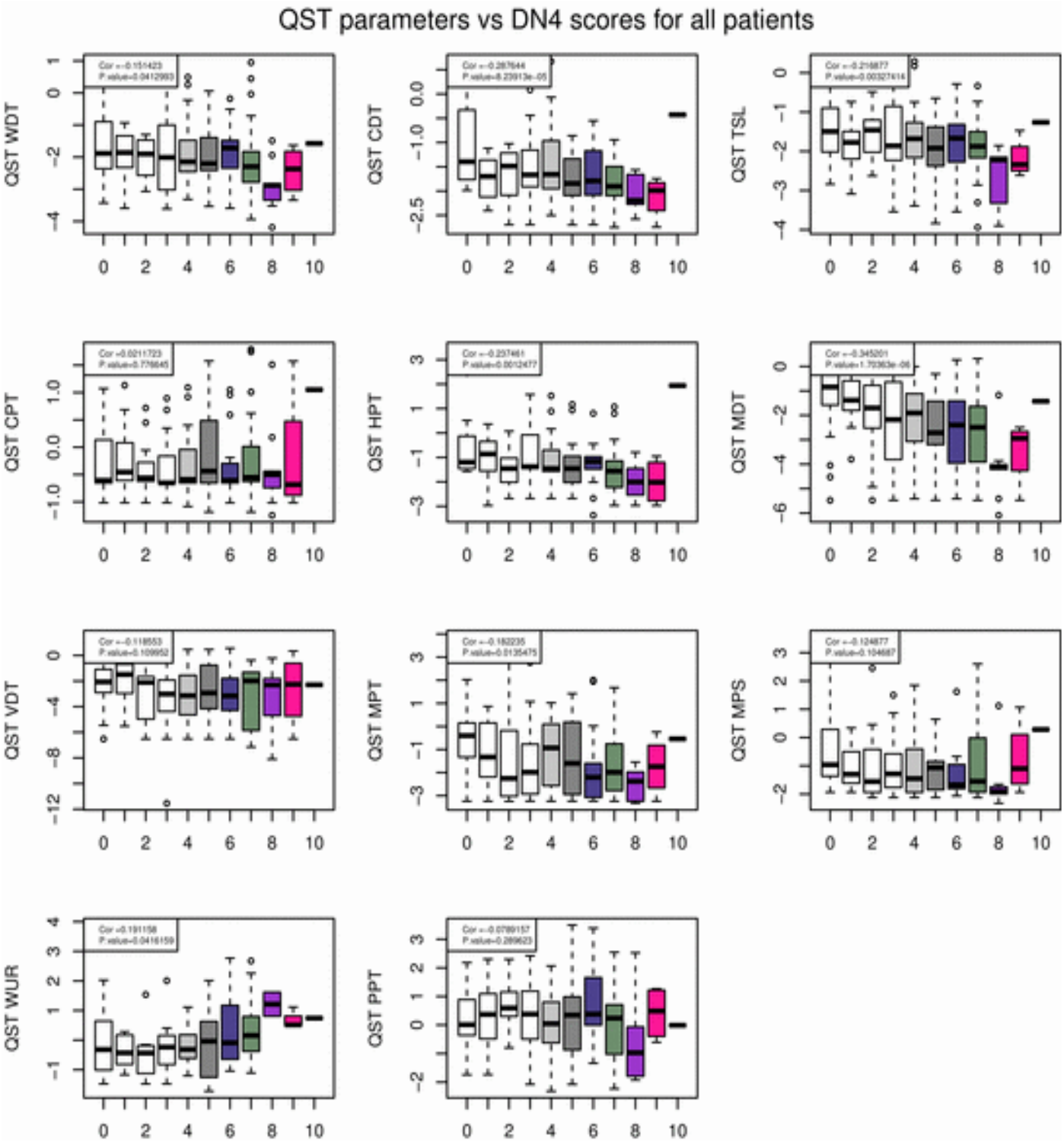


Figure 12: Correlation of DN4 scores to the QST parameters. MDT (correlation coefficient = −0.345201 , p.value = 1.70363e−06) and CDT (correlation coefficient = −0.287644 , p.value = 8.23913e−05) showed significant and strong correlation. TSL, HPT and MPT showed significant moderate correlation.

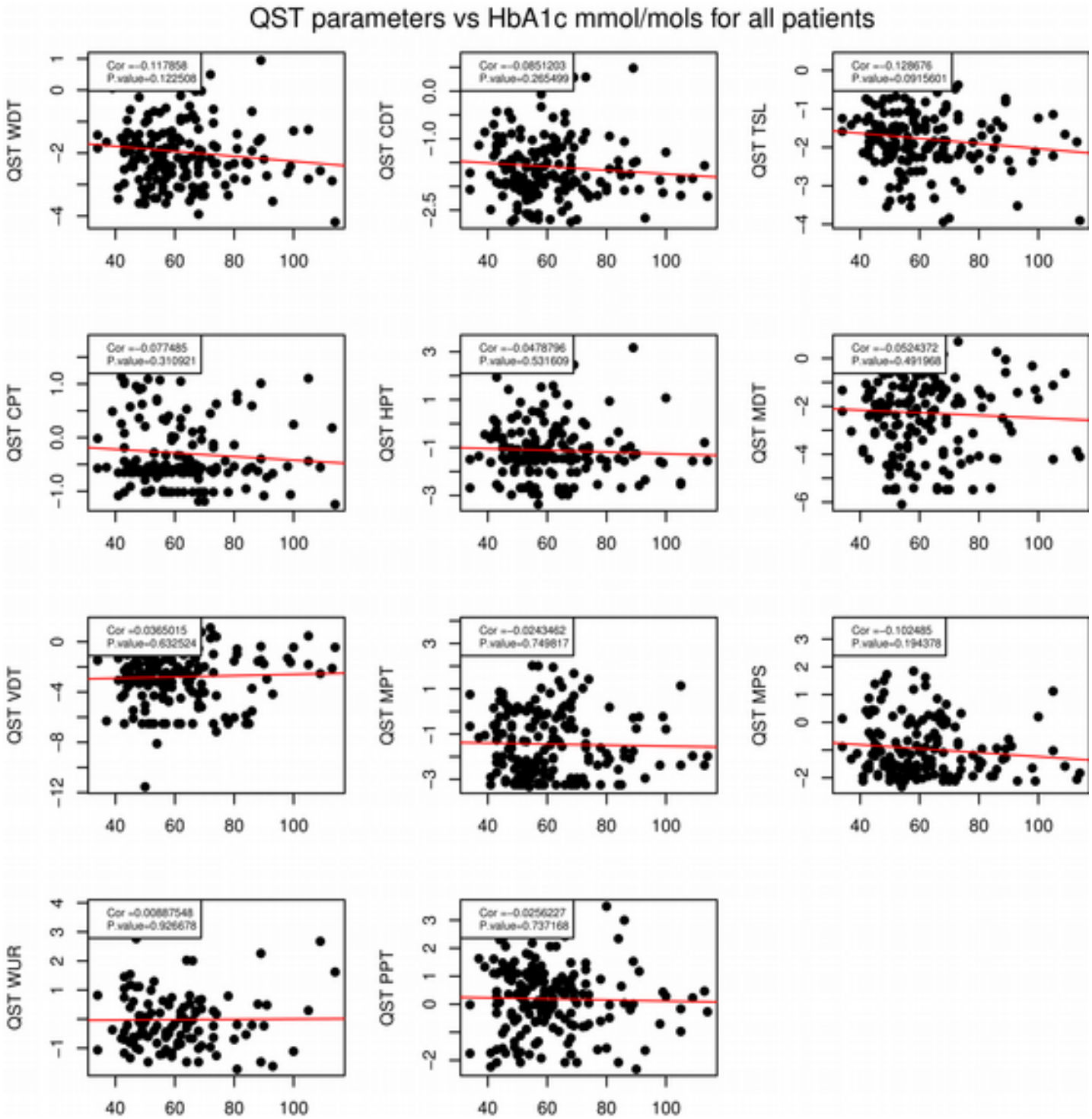


Figure 13: Correlation of HbA1c mmol/mol to the QST parameters. In general we observed weak correlation.

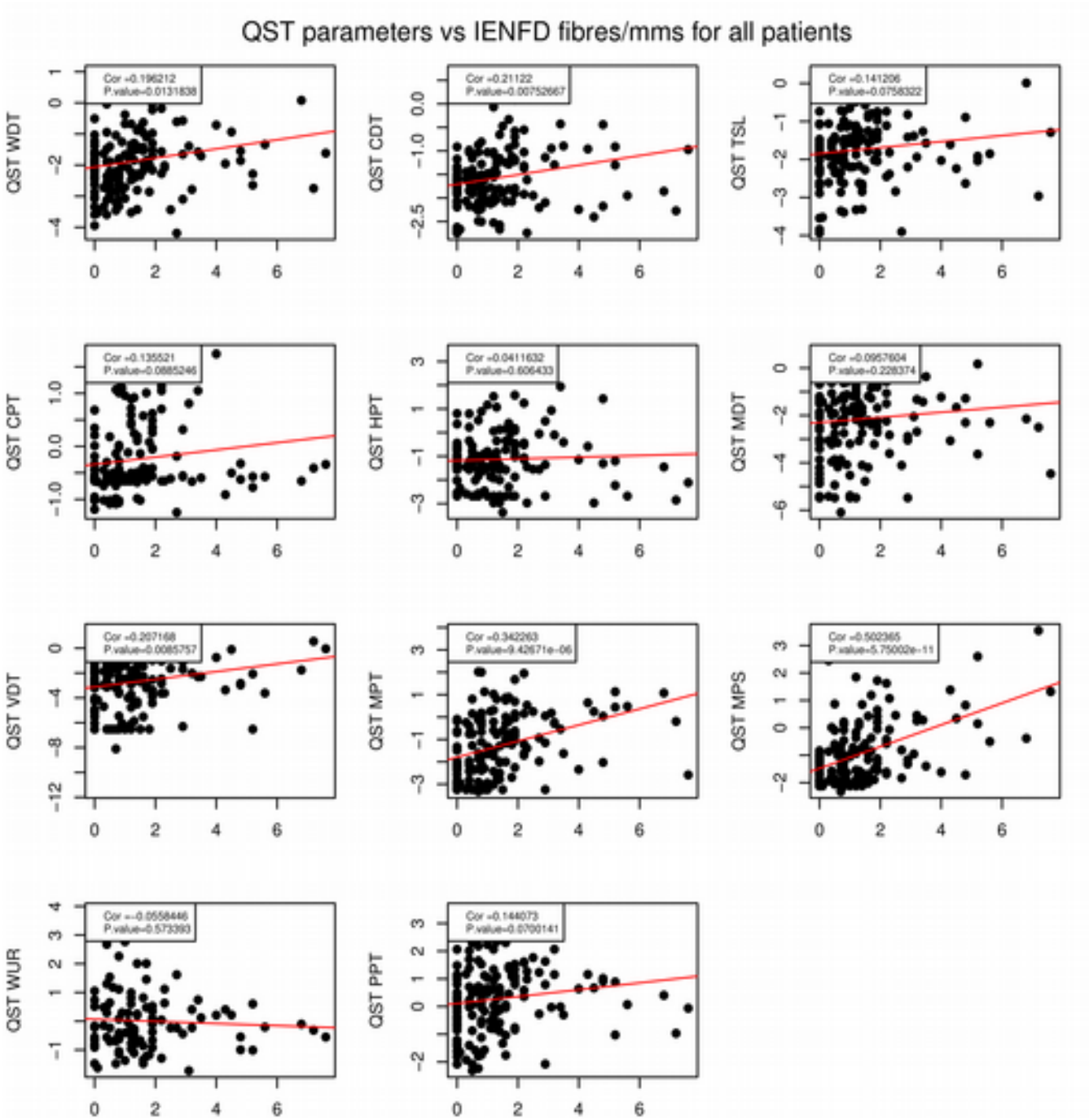


Figure 14: Correlation of IENFD mmol/mol to the QST parameters. Most of the QST parameters were significantly and strongly correlated with IENFD.

Patients with painful neuropathy

Consistent with the above findings we did not find any significant correlations with QST and the HbA1c concentration, but we did find again significant correlations with lower IENFD leading to lower QST thresholds, figure 15. Thus we were able to confirm, also in the group of patients with painful neuropathy, that QST can assess very well sensory deficits related to reduced IENFD, but as expected these deficits are not correlated to pain intensity. The most strongly and significantly associated score was the Mechanical Pain Sensitivity (MPS) with Pearson's correlation coefficient = 0.46463 and p.value = 5.10899e−06. Moreover WDT, CDT and TSL scores related to cold and warm detection threshold and to paresthesia and the MPT, CPT score related to mechanical and cold pain threshold were highly and significantly correlated to IENFD. On the other hand QST is not correlated with the pain intensity as assessed by the NPSI questionnaire, figure 16. This is an expected result as we have already found that NPSI pain intensity is not associated to IENFD but rather to HbA1c.

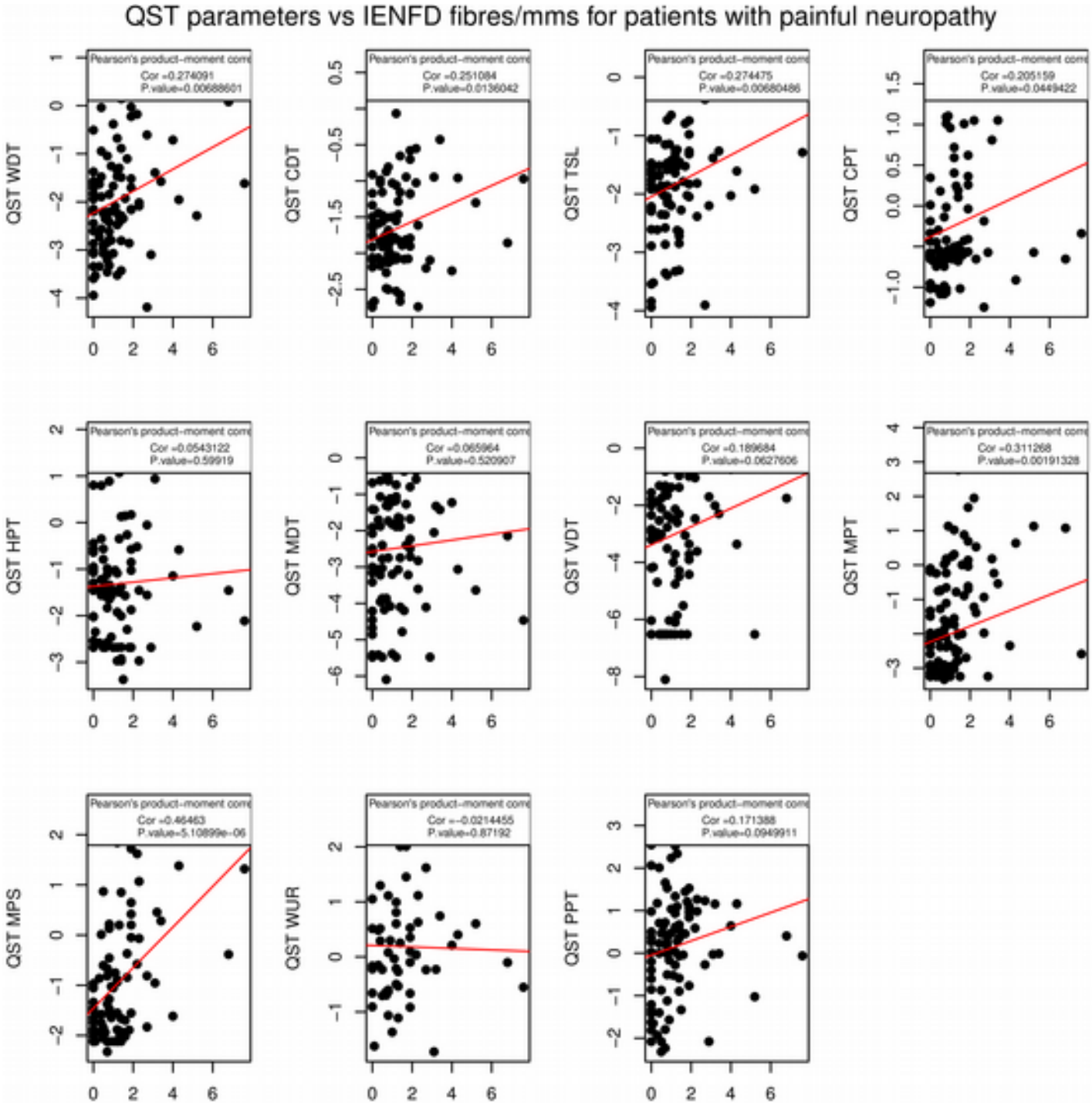


Figure 15: QST parameters are highly correlated to the IENFD

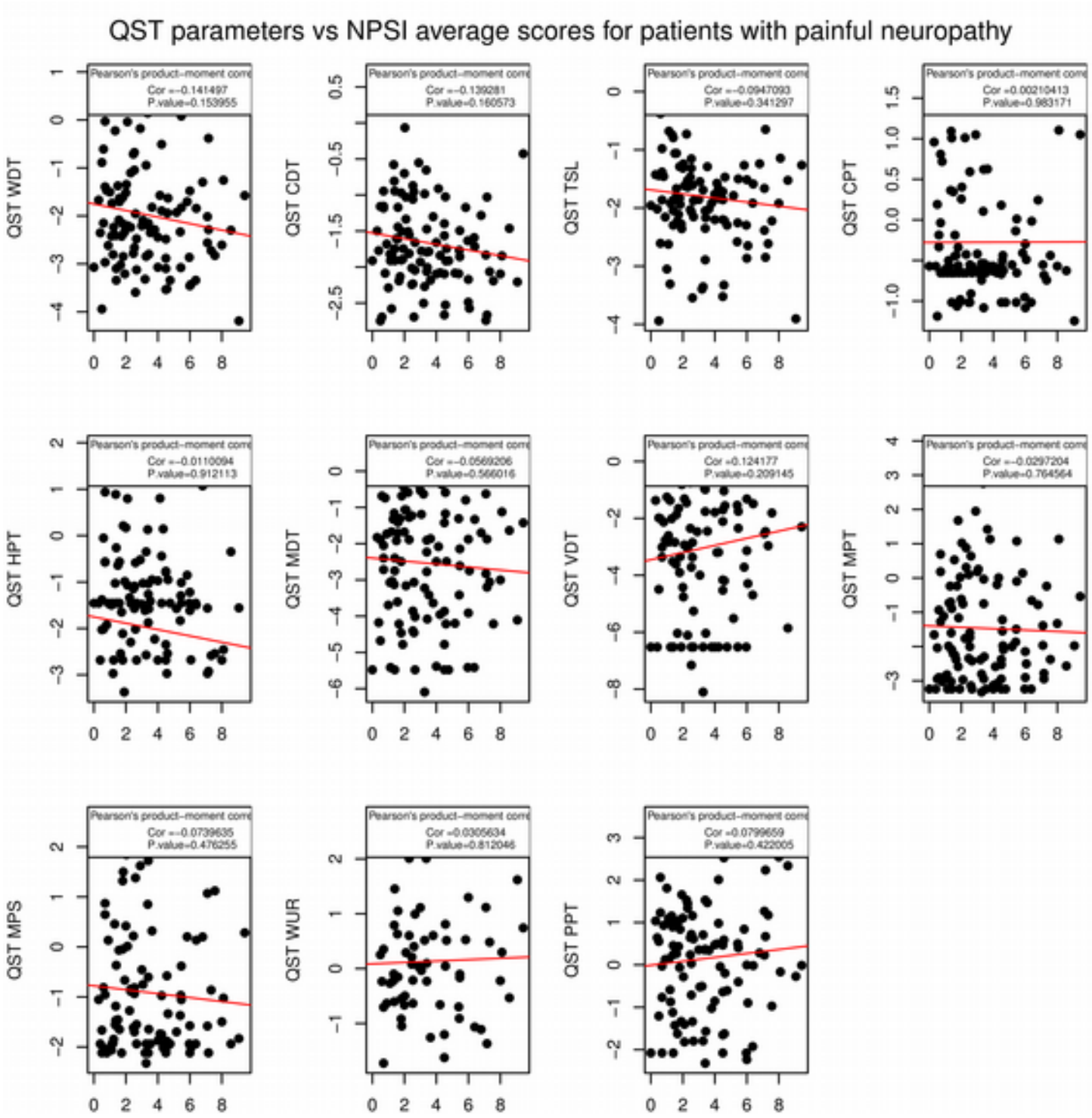


Figure 16: QST parameters are not strongly correlated with NPSI average pain score. For QST scores WDT, CDT, TSL, HPT, MDT, MPS the higher the pain intensity the lower the QST thresholds but we did not observe neither strong nor significant correlation. QST WUR, PPT and VDT had higher values in patients with higher pain intensity.

Principal Components Analysis and clustering

Next we proceeded to principal component analysis of NPSI and QST questionnaires aiming to identify pain dimensions that reflect to the severity of neuropathic pain phenotype and clusters of patients with distinct pain modalities. From the above results we expected that clustering based on NPSI data would better reflect the phenotype and that HbA1c, Gender and Age would be highly associated with clusters of patients with different severity of painful neuropathy.

After normalization and handling of missing values we carried out principal components analysis (PCA). We plotted scree plots and selected the number of principal components after which the rate of the eigenvalue reduction and the rate of rise of the explained variance gets significantly slower, i.e. the elbow in the plot of the number of components vs eigenvalues and the number of components vs cumulative percentage of variance. Then we calculated the loadings which we transformed using the varimax rotation (see Introduction, section Principal Components Analysis).

Consequently we input these varimax rotated components into a k-means clustering algorithm. We used 100 initial random positions of the cluster centroids and allowed the algorithm to converge in 200 iterations. As k-means clustering depends on a predefined number of clusters, *k*, we have used the same method as for principal components, for defining the optimal number of clusters. We have plotted the within clusters sum of squares, i.e. the sum of the squared differences of each observation of a group from the respective group's mean against the number of clusters. Where we observed a distinct drop, i.e. elbow, in within groups sum of squares, when we moved from a solution with a certain number of clusters to another, we identified the best fit.

We also clustered patients according to the values of these varimax rotated components using hierarchical clustering based on the euclidean distance between data points and the ward method for the actual clustering. Then we cut the tree produced by the hierarchical clustering so as to have the same number of groups that optimally partitioned data according to the

within groups sum of squares as described above. We followed the exact same process analysing both QST and NPSI data. QST's principal components failed to separate samples according to pain severity and also clustering did not produce any groups of individuals associated with neuropathic pain severity, figure 17. This is consistent with the previous findings, as QST is correlated mainly with IENFD which is not highly associated with the severity of neuropathic pain. The results from NPSI data analysis are presented in detail below.

After the calculation of the principal components (PC) of NPSI's data we examined how much of the data's variance could be explained by these PCs. For each dataset, namely patients with painful neuropathy using all variables and patients with painful neuropathy using only quantitative variables, we found that we should optimally retain 3 or 4 principal components based on the two elbows in the respective scree plots which cumulatively explain from 66.28% (painful neuropathy only quantitative variables) to 67.72% (painful neuropathy only quantitative variables all variables) of the original data variance, figure 18.

As NPSI is a tool for assessing pain severity and identifying modalities of pain in patients with neuropathic pain we focused only on the dataset of patients with painful neuropathy.

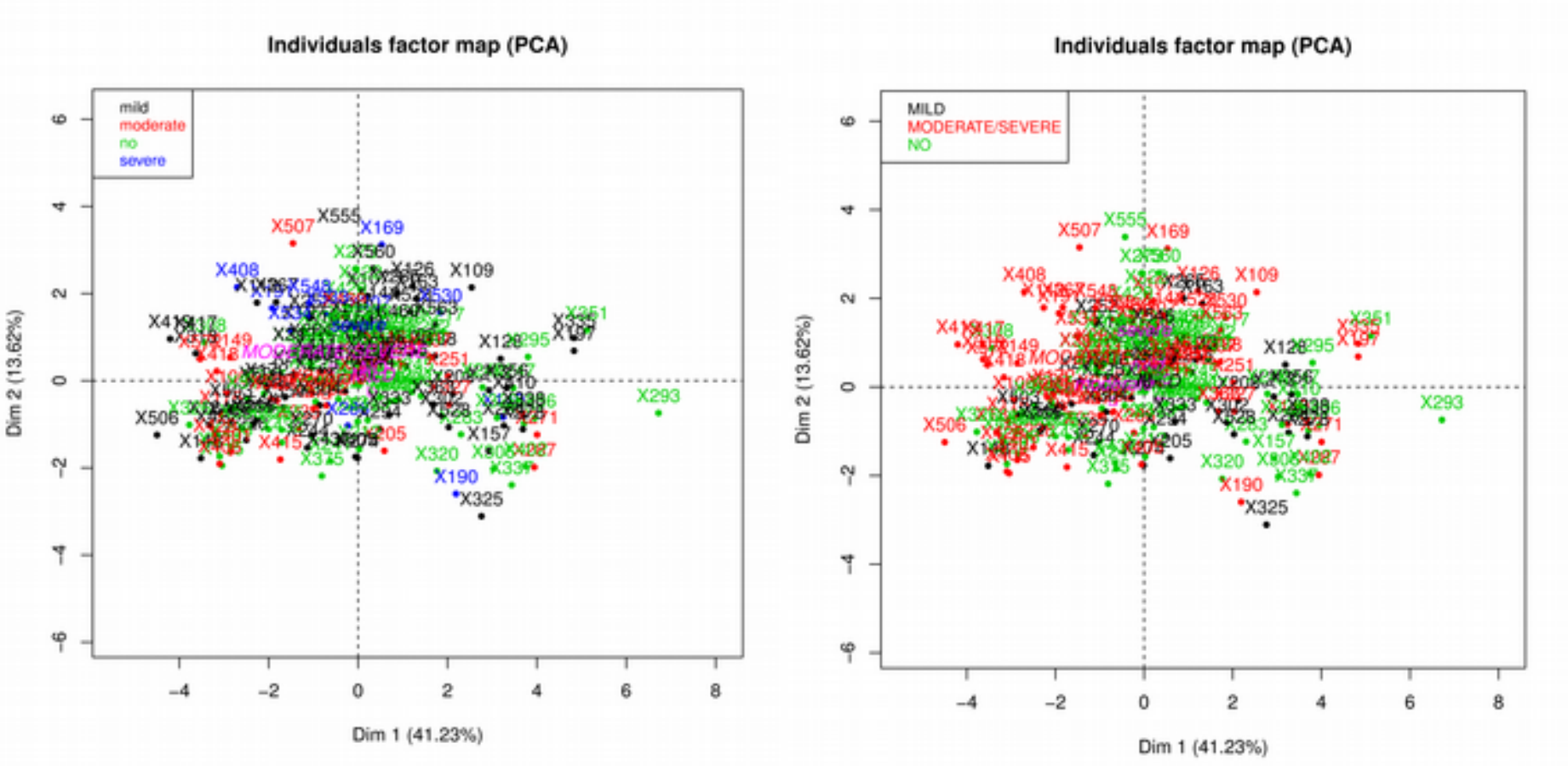
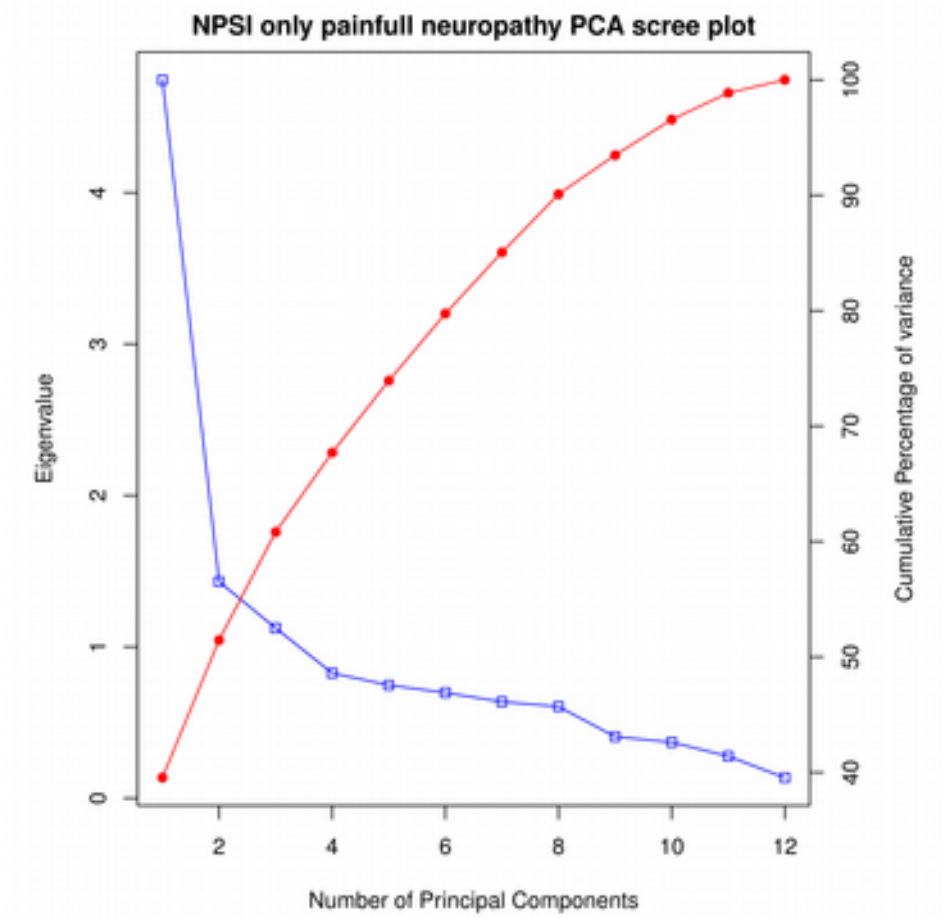


Figure 17: Individuals factor map according to the first two principal components of QST data. No separation of patients according to pain severity assessed, was observed either for the NPSI (left) or 7-day pain diary (right) scores.



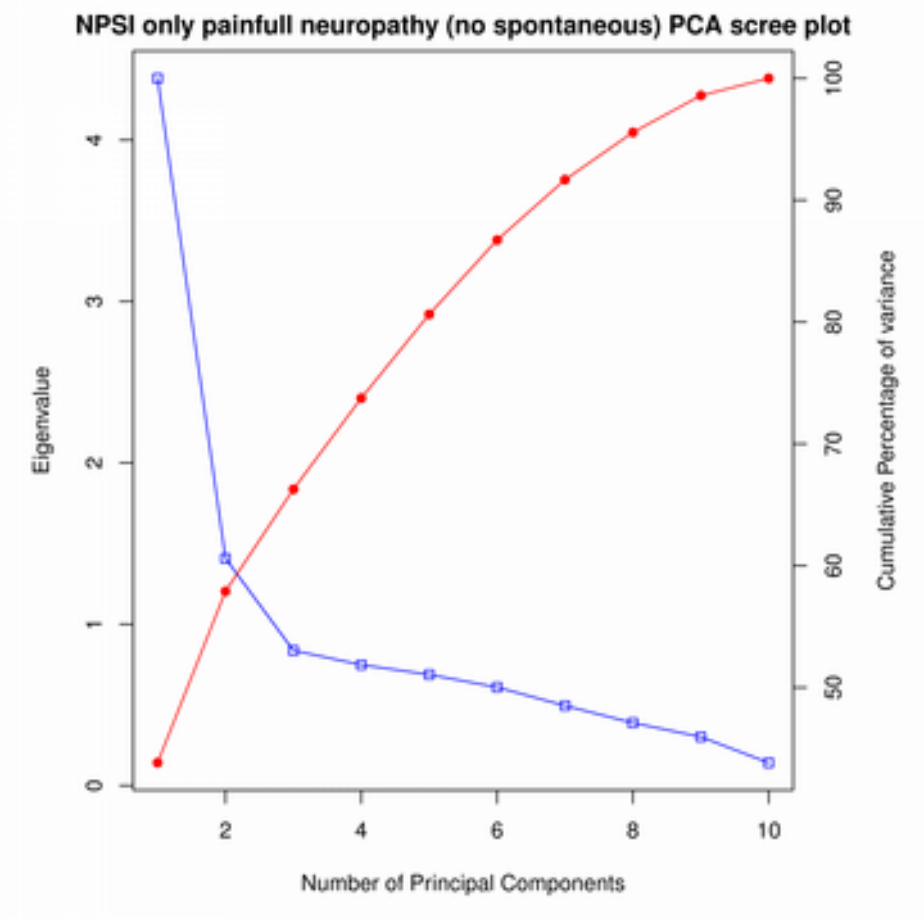


Figure 18: Scree plots for patients with painful neuropathy all variables (left), no spontaneous pain variables (right). Blue line represents the number of principal components vs the eigenvalue, red line represents the number of components vs the cumulative percentage of variance they explain. In the first case the plot's elbow is in the 4th component, in the second in the 3rd.

Factor analysis and contributions to the Principal Components

As our main aim is to understand and identify distinct qualities of neuropathic pain it is of great importance to study how different variables / factors of the NPSI questionnaire contributed to each of the principal components. To do so we calculated the loadings for each PC. Regarding this we observed that different groups of NPSI variables, were consistently co-occurring as the highest contributing factors for certain dimensions / principal components of NPSI data. We hypothesized that these groups represent certain pain modalities.

By observing the variables that are usually better explained together in the same PC we can identify variables which cumulatively define distinct qualities of pain. Moreover as we used the varimax rotation to rotate the orthogonal basis of the loadings of the top components we observed a much better separation of pain modalities, as there was a certain group of factors with high values for each component, figure 19.

When we considered the grouping of individual variables by the principal component they contributed to primarily, we identified 3 or 4 distinct pain modalities, depending on the inclusion of the spontaneous paroxysmal and ongoing pain variables. These distinct pain modalities are very consistent as the grouping remained the same when we included the two categorical variables and also have some similarities with the findings of (Freeman et al., 2014). Interestingly we can describe these pain qualities as Superficial pain and paresthesia/dysesthesia which is comprised by the three relevant NPSI variables and contributed mostly to the 1st component or 2nd component; Evoked pain which is comprised by Brush Evoked, Pressure Evoked, Cold Evoked and Pressure variables and contributed mostly to the 2nd component or to the 1st component when we included the spontaneous pain variables; Deep and Spontaneous pain contributed most to the 3rd component or to the 4th component when we included the spontaneous pain variables; and Spontaneous Paroxysmal pain measured from the two categorical variables which contributed mostly to the 3rd component when we included them in our analysis.

Moreover we observed spatial arrangements that were related to the average pain severity calculated from the NPSI pain questionnaire when we plotted data using the first two principal components for each individual, figure 20. Also all principal components were highly correlated with the NPSI average score and individual patients could be separated according to pain severity based on the first two principal components. Moderate pain had the most widespread spatial localization while patients with mild and severe pain are highly localised showing less variance in the values of their first two principal components.

NPSI varimax components

	Dim.1	Dim.2	Dim.3
NPSI_BURNING	**0.3874855044**	0.1402762788	-0.0968122406
NPSI_SQUEEZING	-0.0185081649	0.30007923	**0.3108276521**
NPSI_PRESSURE	-0.0528587921	**0.4794883509**	0.1475736514
NPSI_ELECSHOCKS	**0.3891324328**	-0.1203497827	0.3667876315
NPSI_STABBING	-0.0489514459	0.0604074278	**0.7899390011**
NPSI_BRUSHEVOKED	0.0346708561	**0.4778519953**	-0.0661799394
NPSI_PRESSUREEVOKED	-0.0137815074	**0.4906509384**	0.0212559341
NPSI_COLDEVOKED	0.1549155265	**0.4079293421**	-0.3290505729
NPSI_PANDN	**0.6050216663**	-0.0819592014	-0.0158411704
NPSI_TINGLING	**0.5490501346**	0.0053182329	0.0184922917

NPSI varimax components – no categorical variables for spontaneous pain

	Dim.1	Dim.2	Dim.3	Dim.4
NPSI_BURNING	0.1124899719	**-0.359220715**	0.1032697652	-0.085339525
NPSI_SQUEEZING	0.3049677064	0.0174596705	-0.0241409905	**0.3167345465**
NPSI_PRESSURE	**0.4411794206**	0.0758380247	0.1640888656	0.1530656024
NPSI_SPONTONGOING	0.007025064	0.0944727571	**0.7749909408**	-0.037758459
NPSI_ELECSHOCKS	-0.1296308841	**-0.3850888993**	0.0307800766	0.3562147282
NPSI_STABBING	0.0682984314	0.0343873183	-0.0540317709	**0.7848606766**
NPSI_SPONTPAROXYSMAL	0.0105924115	-0.1480684176	**0.5729137354**	0.0065411848
NPSI_BRUSHEVOKED	**0.4999317558**	-0.022279447	-0.0719370702	-0.04227831
NPSI_PRESSUREEVOKED	**0.4789664156**	0.0102039173	0.0346699114	0.0215214762
NPSI_COLDEVOKED	**0.4370140745**	-0.192682759	-0.1321704758	-0.349002015
NPSI_PANDN	-0.0862476711	**-0.5848946933**	0.0250498181	-0.014716396
NPSI_TINGLING	0.0176204664	**-0.5520416507**	-0.0681614815	0.0182757593

Figure 19: 3 or 4 pain dimensions were identified by the principal component analysis. Varimax rotated components had higher contributions of certain NPSI variables. We have coloured NPSI variables according to which PC they contributed to most. Thus a profile of Superficial pain and Paresthesia/dysesthesia (red) has higher values in Dim.1 or Dim.2; Deep and spontaneous pain (green) has higher values in Dim.3 or Dim.4; Evoked pain (yellow) in Dim.2 or Dim.1; Spontaneous paroxysmal pain (blue) in Dim.3. Non varimax rotated components are in Appendix 5.

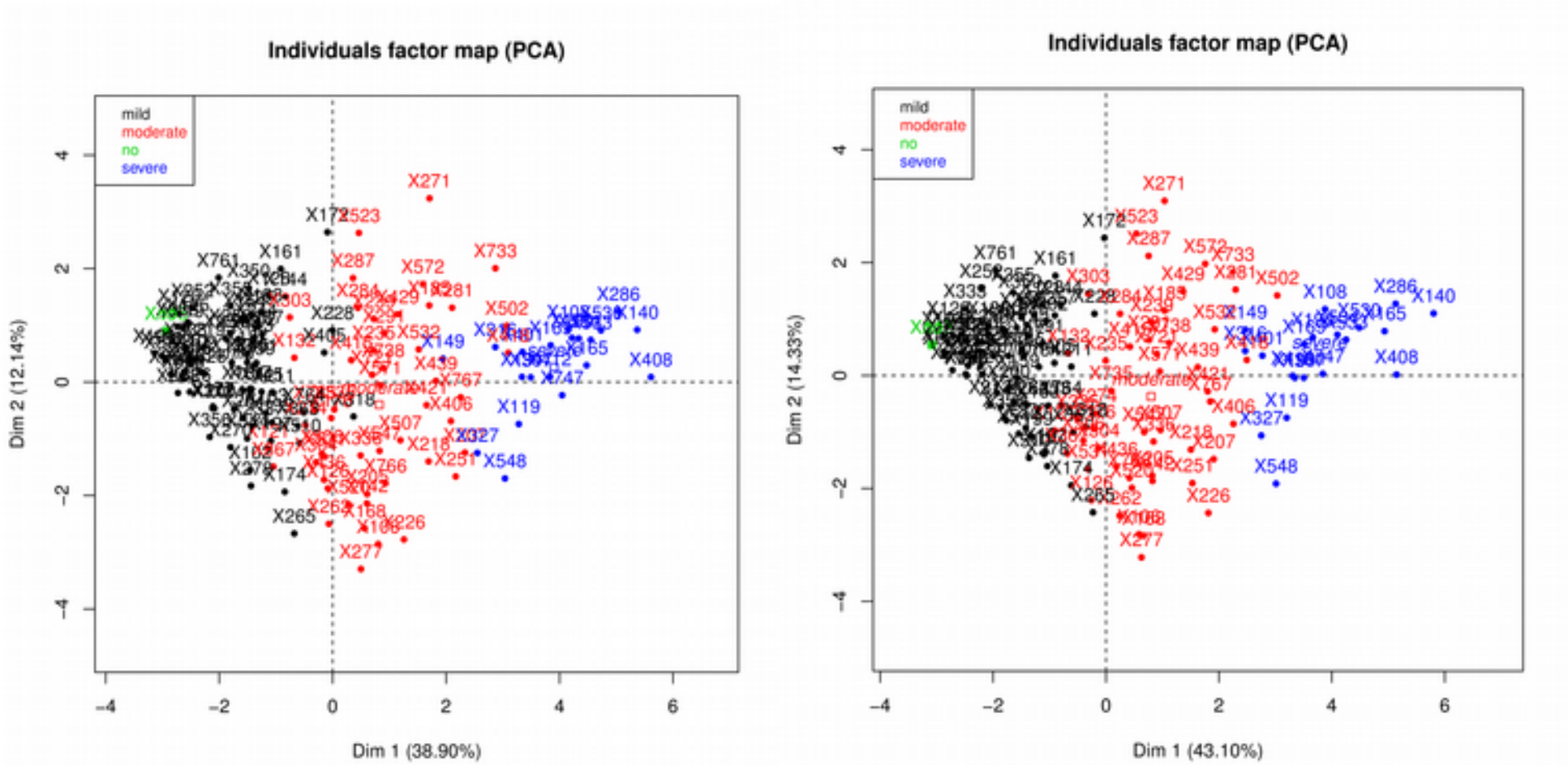


Figure 20: Individuals factor map (PCA plot) for painful neuropathy top left and for painful neuropathy without spontaneous pain variables top right. Severe average NPSI scores are blue, moderate are red and mild are black. Individuals can be separated according to pain severity from these two first principal components.

NPSI principal components were correlated to clinical markers

After we calculated the NPSI principal components we examined them in the context of the clinical variables that could be associated with them. First, as expected given the NPSI score's high correlation, we have significant correlation of these components to the age and gender. We present figures from the dataset of patients with painful neuropathy which include all variables, but results were almost the same for the three retained components when we excluded the Spontaneous paroxysmal pain categorical variables. Principal components were significantly negatively correlated to age, whereas the component of superficial pain and paresthesia/dysesthesia showed higher and most significant negative correlation to age, figure 21.

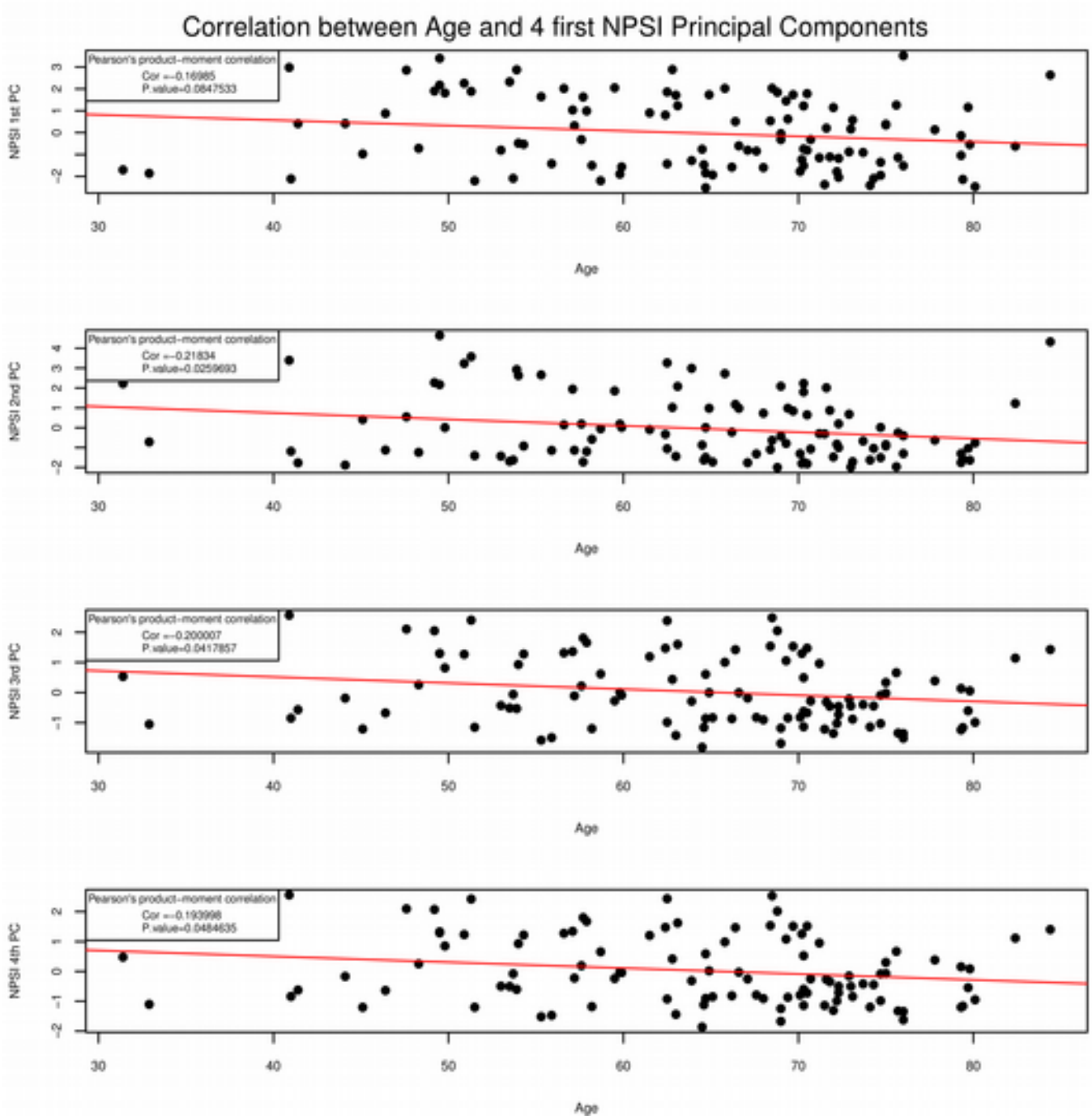


Figure 21: All PC were negatively correlated to age. The 2nd component, corresponding to superficial pain and paresthesia/dysesthesia showed slightly stronger negative correlation

Additionally there is a strong effect of the patients gender in these principal components. Females consistently reported more intense pain and showed higher values in principal components as well, figure 22. Regarding clinical parameters there was no significant correlation between PCs and the

duration of diabetes or BMI, only weak and non-significant negative correlation between PCs and IENFD and strong and significant correlation between HbA1c and the varimax rotated PCs, figure 23.

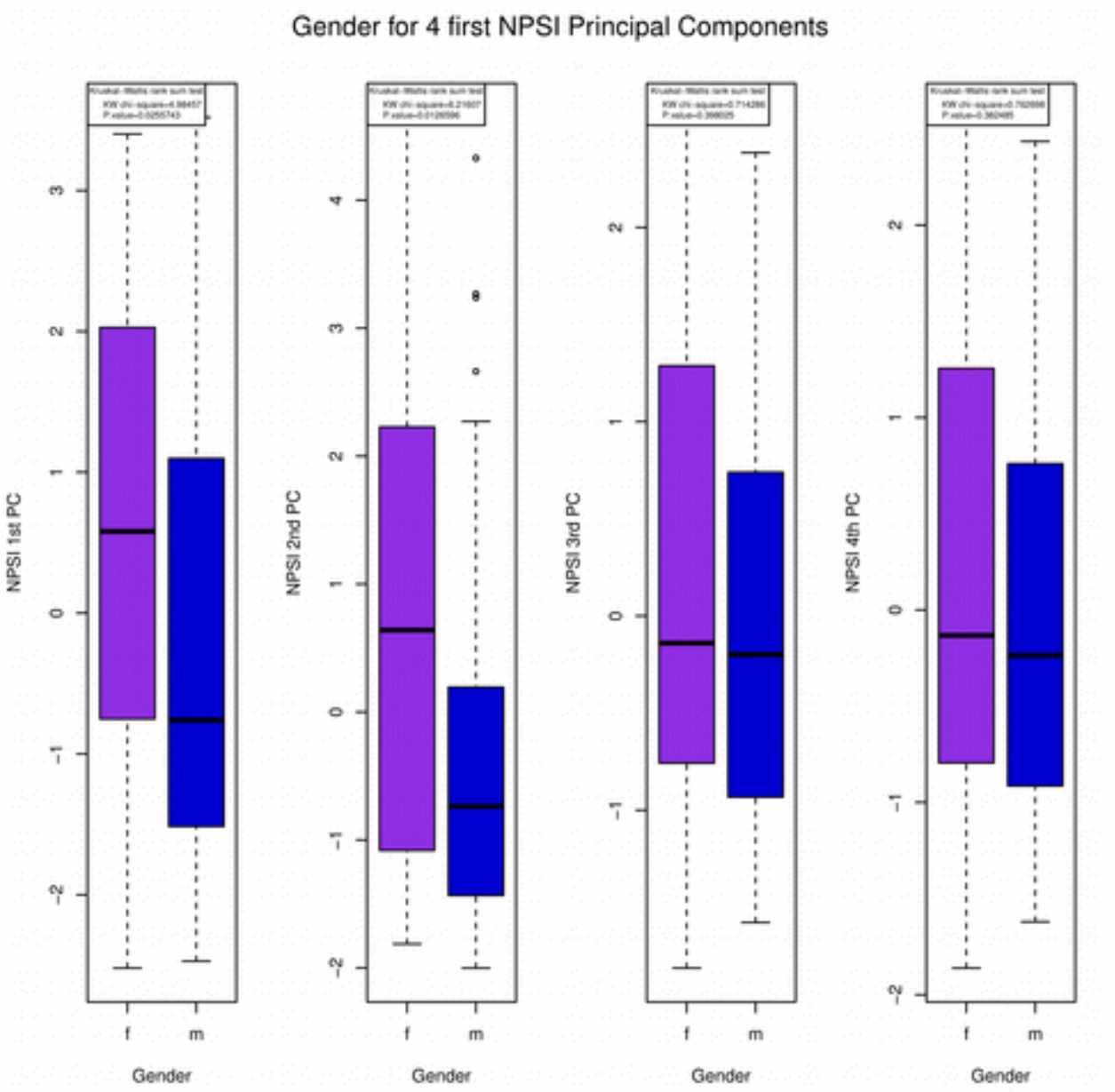


Figure 22: Females (violet) had higher values in PCs than males (blue)

In this context, the 3rd component which contributes higher in spontaneous paroxysmal and ongoing pain has the stronger and more significant positive correlation to HbA1c concentration.

The above principal component analysis identified distinct qualities of pain in the dataset. Moreover, we demonstrated that these pain dimensions are related to pain severity as measured from the average total NPSI score. In figure 20 we visualise how these PCs can effectively separate patients according to pain severity and we also show that they are significantly correlated to age, gender and HbA1c for each patient. Therefore, we hypothesize that we can identify distinct clusters of patients given these varimax rotated components.

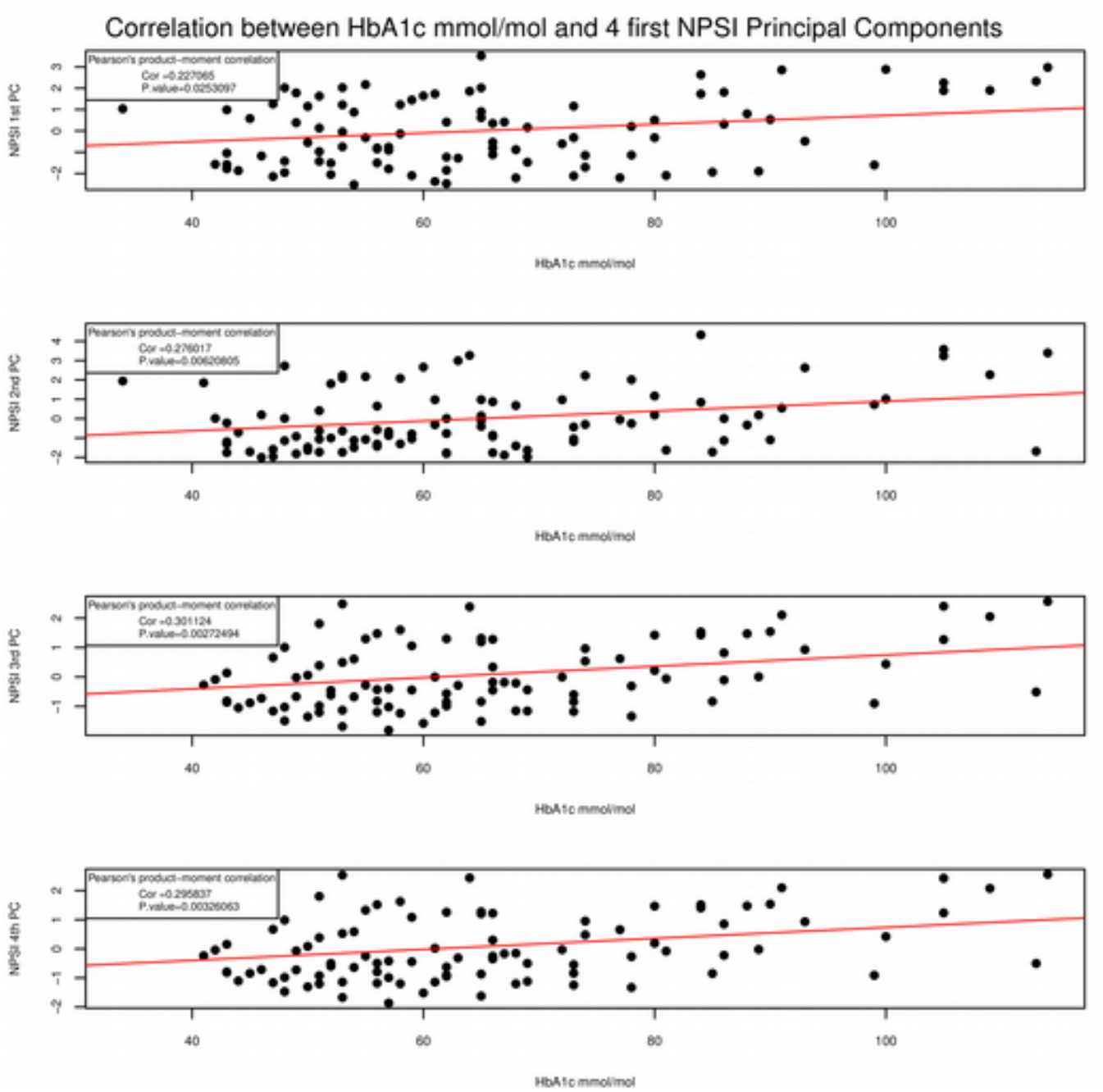


Figure 23: HbA1c mmol/mol concentration is highly correlated to the values of the 4 principal components. The 3rd component, of spontaneous paroxysmal and spontaneous ongoing pain showed the higher and most significant correlation.

Clustering

In order to identify clusters of patients we input these varimax rotated principal components into the centroid and hierarchical clustering algorithms. More specifically we carried out k-means clustering and euclidean distance clustering with the Ward criterion (Ward, 1963).

As mentioned above, prior to clustering we determined the number of clusters which produced an optimal separation of data. This is particularly important for the k-means clustering algorithm. To do this we plotted the within groups sum of squares against the number of clusters. A distinct drop to the rate of reducing the within groups sum of squares shows that a solution of 4 clusters of patients might be the optimal number for all datasets (figure 24). Thus this would be the number of centroids for the k-means clustering and the number of groups for the hierarchical clustering.

We then performed k-means clustering and plotted the groups with the pain severity of each individual colour coded. The algorithm started from 100 initial random centroids and then refined the cluster assignments by controlling the within cluster variance for a maximum of 200 iterations. We then colour coded individuals according to pain severity and observed that a distinct cluster was comprised of patients with only mild neuropathic pain, two mixed clusters one predominantly with moderate and one predominantly with mild neuropathic pain and one cluster of patients most of whom were suffering from severe neuropathic pain, figure 24. Thus we were able to efficiently separate patients according to pain severity in an unsupervised way using only 3 or 4 varimax rotated principal components from the NPSI data, figure 26.

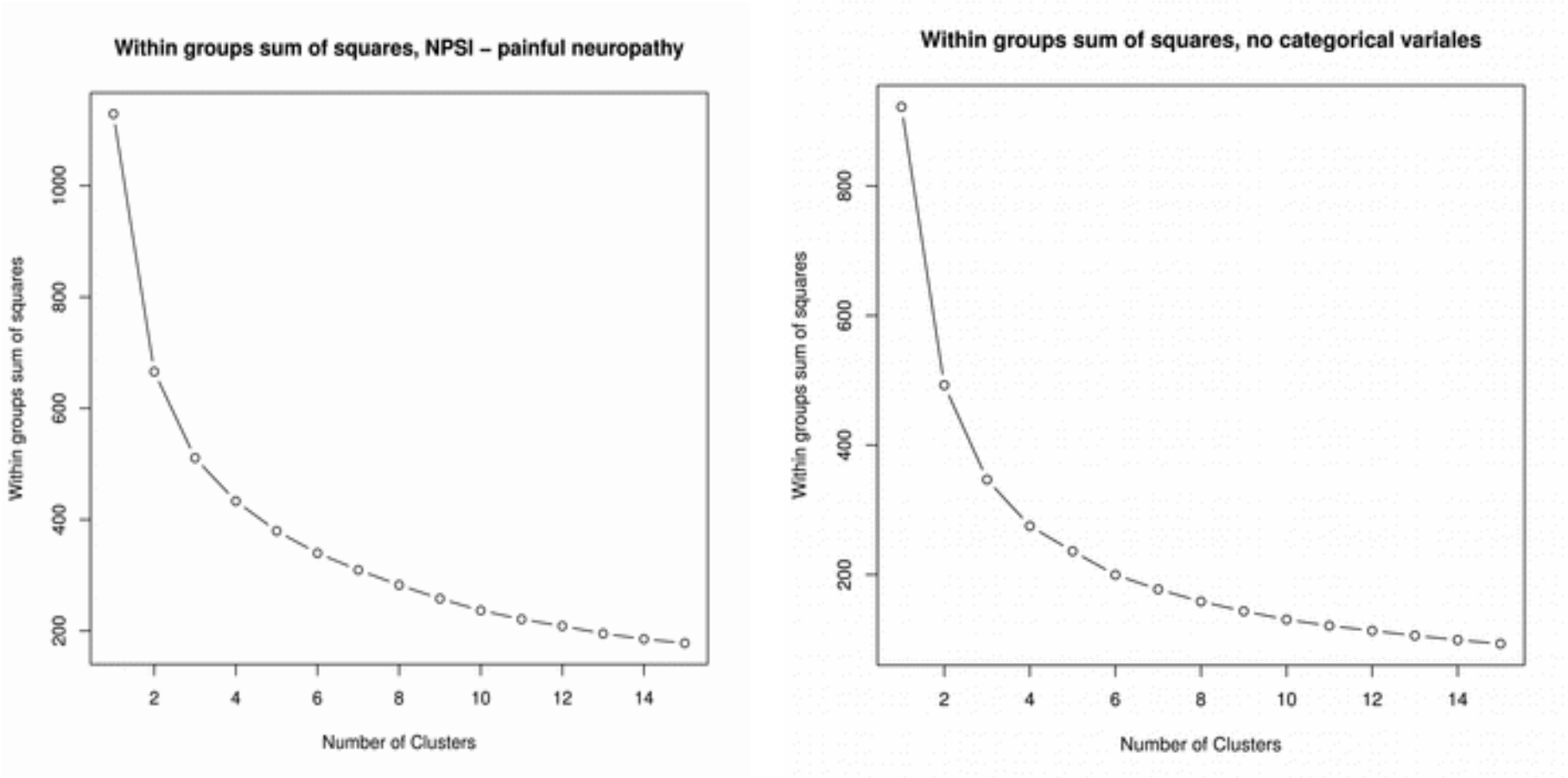


Figure 24: Within groups sum of squares against the number of clusters. Patients with painful neuropathy – all variables (left), patients with painful neuropathy – spontaneous pain categorical variables

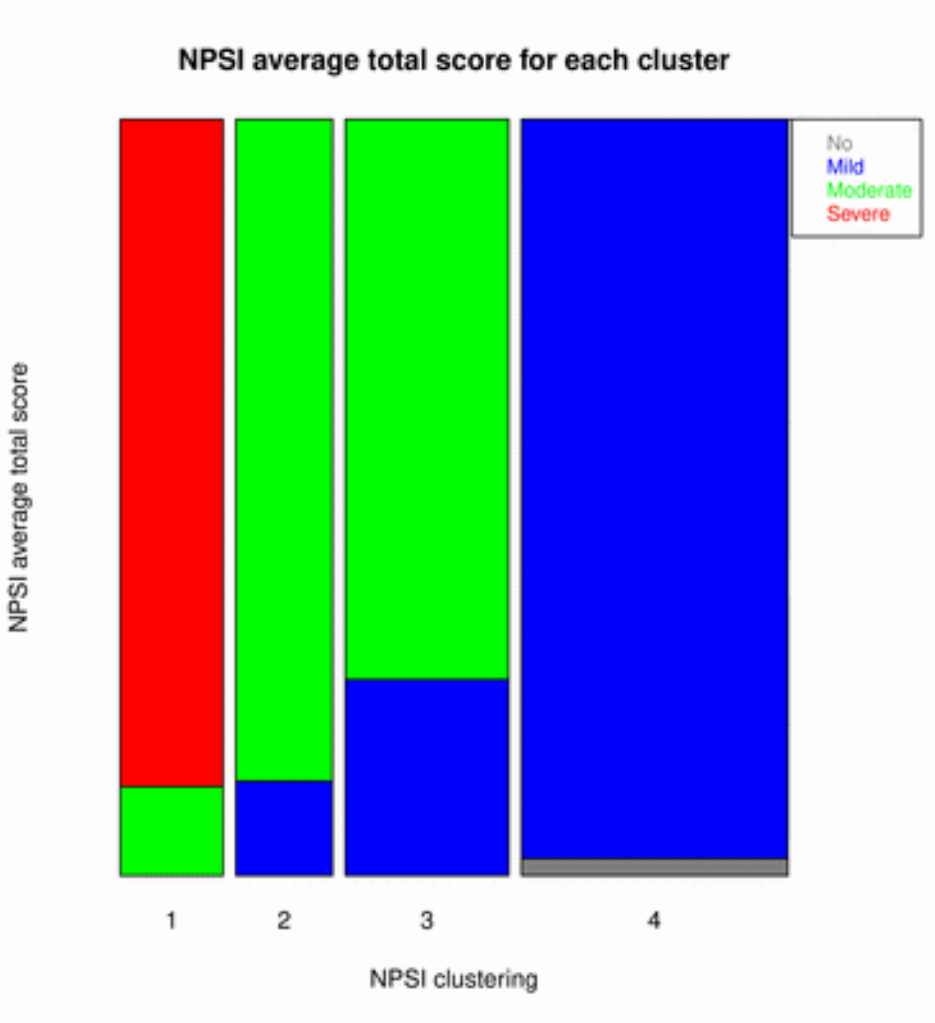


Figure 25: Spineplot of the K-means clustering based on the 3 first varimax rotated NPSI principal components. The 1st cluster has patients with moderate and severe pain. The 2nd and 3rd with moderate and mild. The 4th cluster has only patients with mild pain. Width of the columns represents the cluster's size.

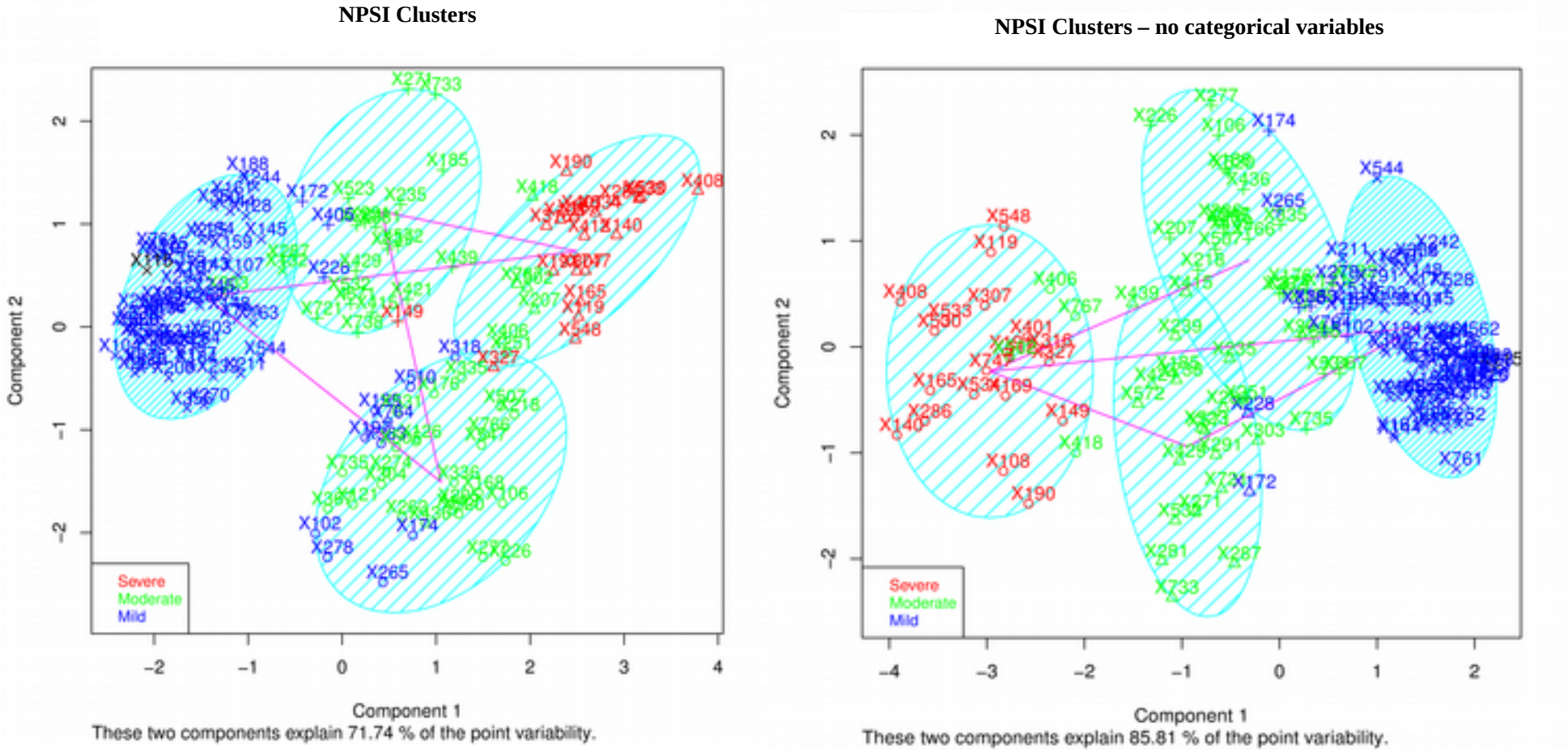


Figure 26: Cluster plots colour coded by pain severity. Left: patients with painful neuropathy – all variables. Right: patients with painful neuropathy – no spontaneous pain variables.

Hierarchical clustering also gave the same results. It separated patients according to pain severity into 4 distinct groups. One with patients having only mild neuropathic pain, one with only severe and some moderate pain and two with moderate and mild pain, figure 27.

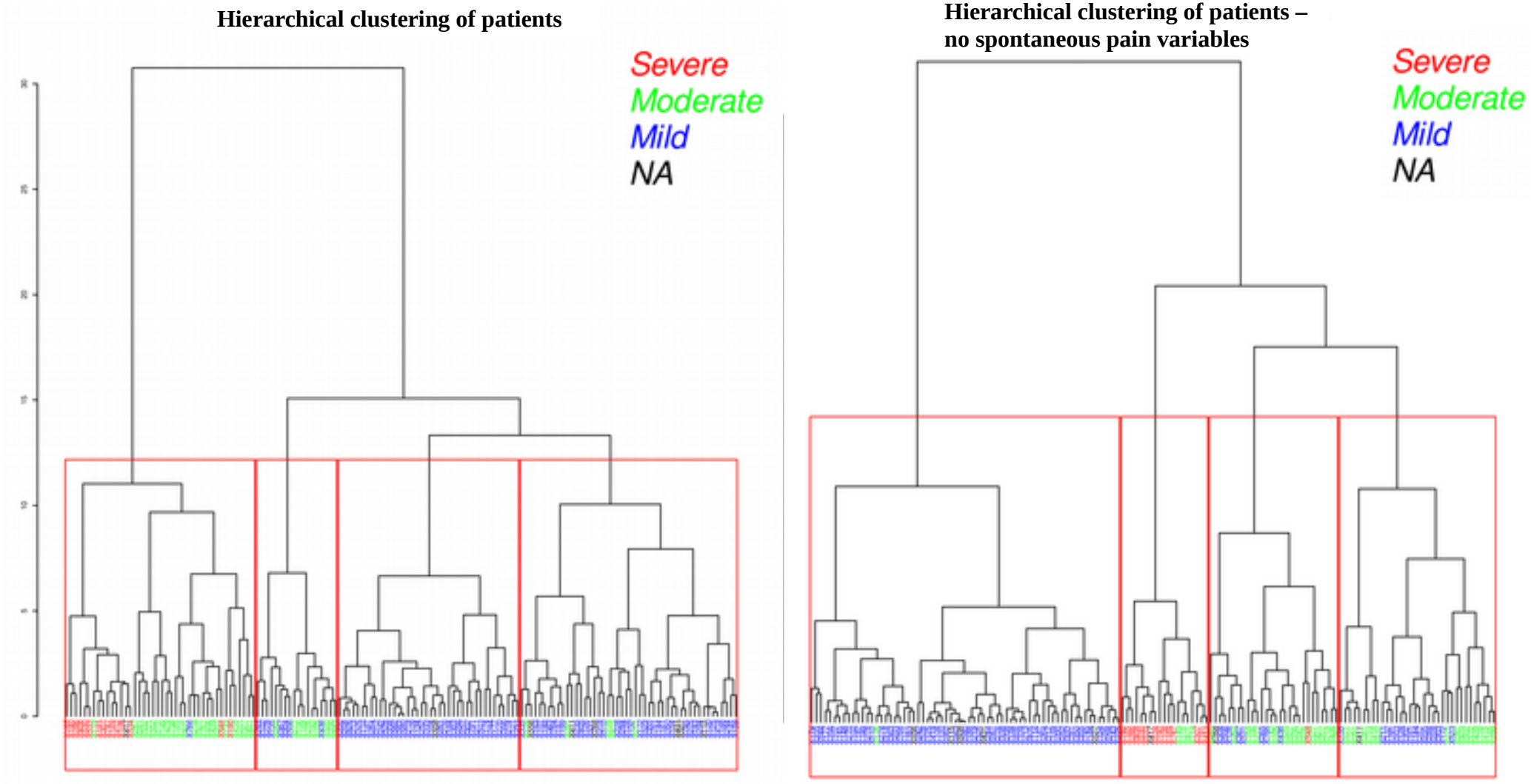


Figure 27: Hierarchical clustering of patients with painful neuropathy (left) and painful neuropathy – no spontaneous pain variables (right)

Association of clusters and principal components to clinical markers

Next we examined how the values of clinical markers were distributed in these 4 clusters of individuals related to pain severity. We first confirmed that clusters were highly associated, not only with NPSI assessed pain severity, but also with the seven day pain diary mean score (Kruskal-Walis test of independence p.value = 5.47873e−11) and with the DN4 score (Kruskal-Walis test of independence p.value = 5.64339e−05), figure 28. Indeed patients assigned to the first cluster (dark red) have consistently higher scores in quality of life pain-questionnaires indicating more severe neuropathic pain. The least severe / mild pain was observed in the 4th cluster (violet), then we have moderate pain in the 3rd and 2nd cluster (the pain diary score separated clusters better than DN4) and severe pain in the 1st cluster (dark red). There was also significant dependence, (Kruskal-walis test of independence p.value =0.0436883) between TCSS symptom sub-score and cluster assignment.

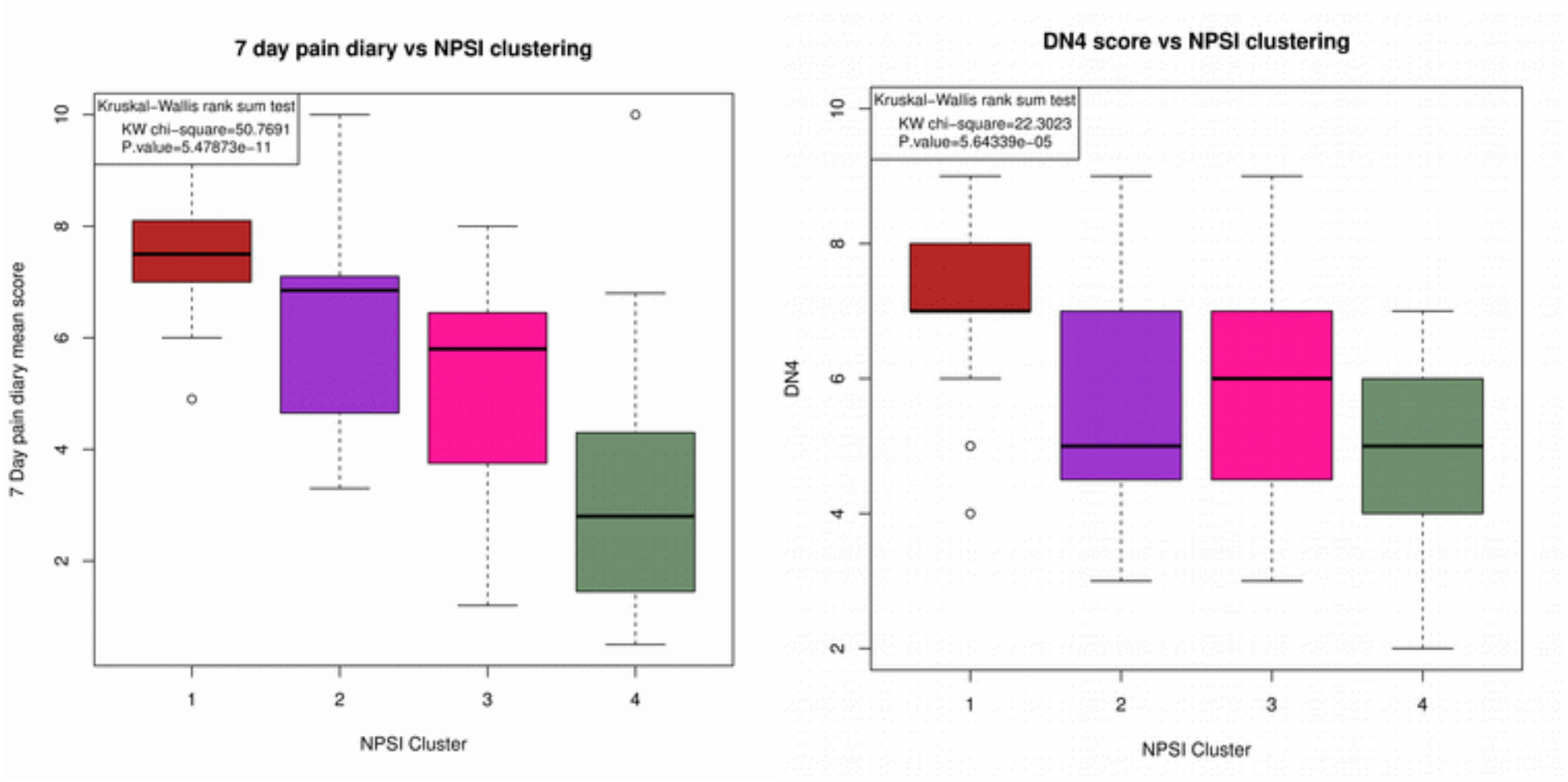


Figure 28: Association between Pain Diary and DN4 scores to NPSI clusters. Both scores are highly associated to cluster assignment. Note that, as expected, all patients with painful neuropathy had DN4 scores > 4.

Since we found that cluster assignment was associated with scores that reflect the neuropathic pain phenotype intensity, we proceeded to examine clinical variables in the context of these clusters.

Consistent with the results from the analysis of QST data, both for the compendium of all patients and for patients with painful neuropathy only, we did not find any significant association (One-way ANOVA test) between QST parameters and cluster assignment. Moreover, as expected there was no significant effect of IENFD on cluster assignment, figure 29, but patients who were grouped together in the 1st cluster of severe neuropathic pain showed higher median values of HbA1c concentration and much higher interquantile range, although high variance of the data did not let the non-parametric Kruskal-Wallis test reach significance.

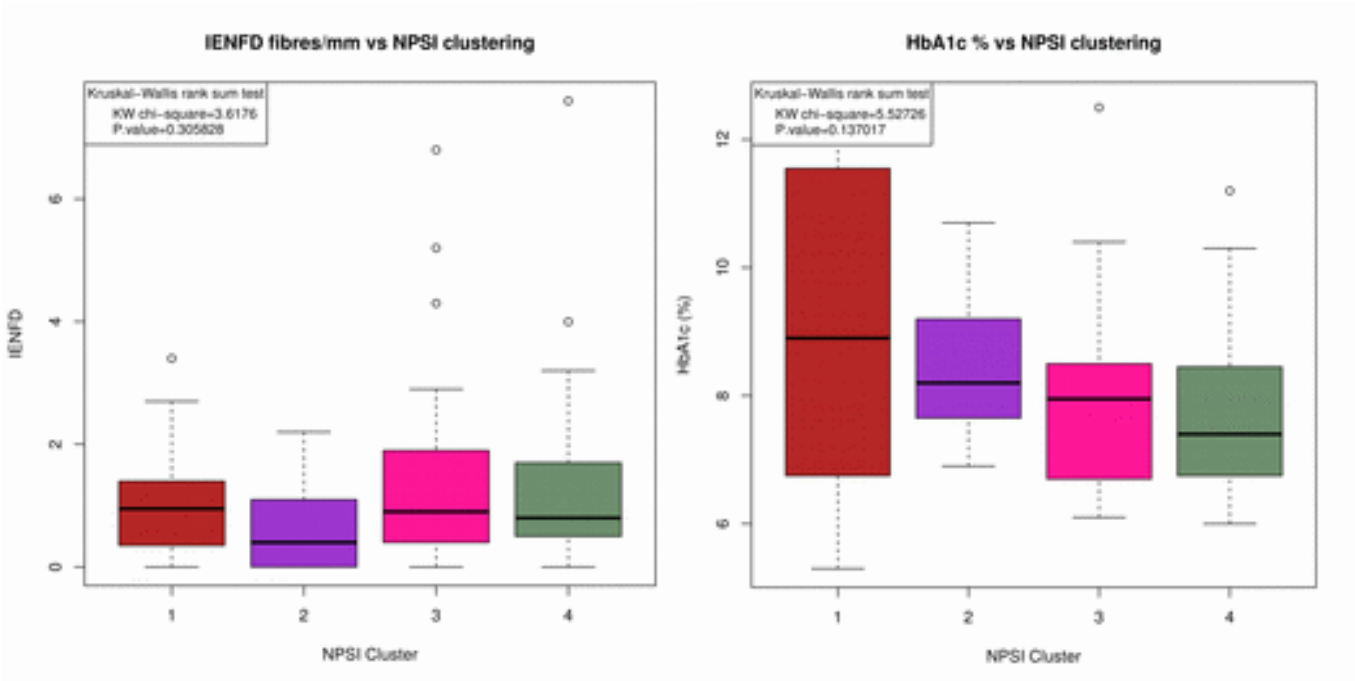
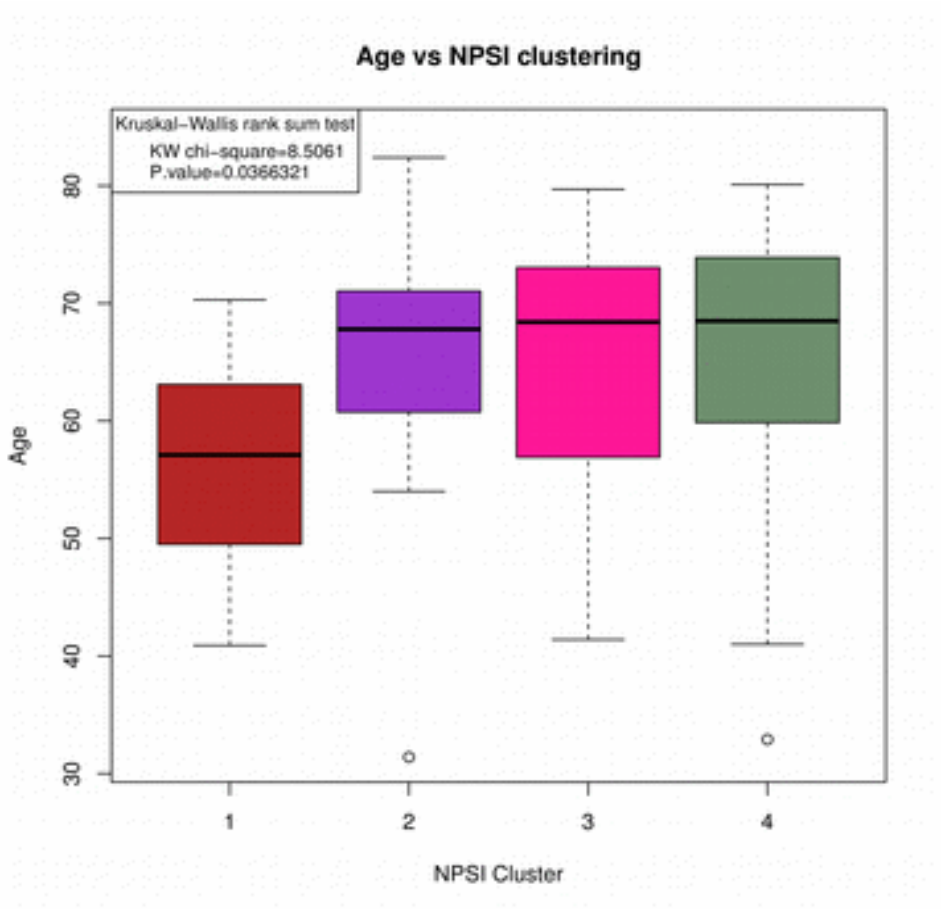


Figure 29: No effect of IENFD on cluster assignment. Patients assigned to a cluster of higher severity have higher HbA1c blood concentrations but the effect does not reach significance.

Moreover BMI and duration of diabetes did not have any effect on cluster assignment, consistent with the finding that they did not correlate to neuropathic pain intensity. On the other hand age and gender were highly associated with the NPSI clustering, figure 30, in a way that younger

patients and females were assigned to clusters of higher pain intensity. Additionally, the TCSS symptoms subscore had a significant effect on clustering, indicating that the NPSI questionnaire, and consequently clusters based on this data, can distinguish patients according to symptoms of painful neuropathy, figure 31. In other words patients grouped in different clusters are not likely to come from populations with similarly distributed TCSS symptom scores.

Finally even though clusters 2 and 3 (violet and pink) were similar in terms of average pain intensity they have significant differences in the distinct pain modalities observed in them. We should note here that Pain Diary (figure 28) revealed a fine grading of pain intensity between clusters but other scores, including NPSI, and clinical markers, showed similar median values. Interestingly these clusters grouped patients according to subtle differences in pain modalities that cannot be observed in the general NPSI total score or its average value.



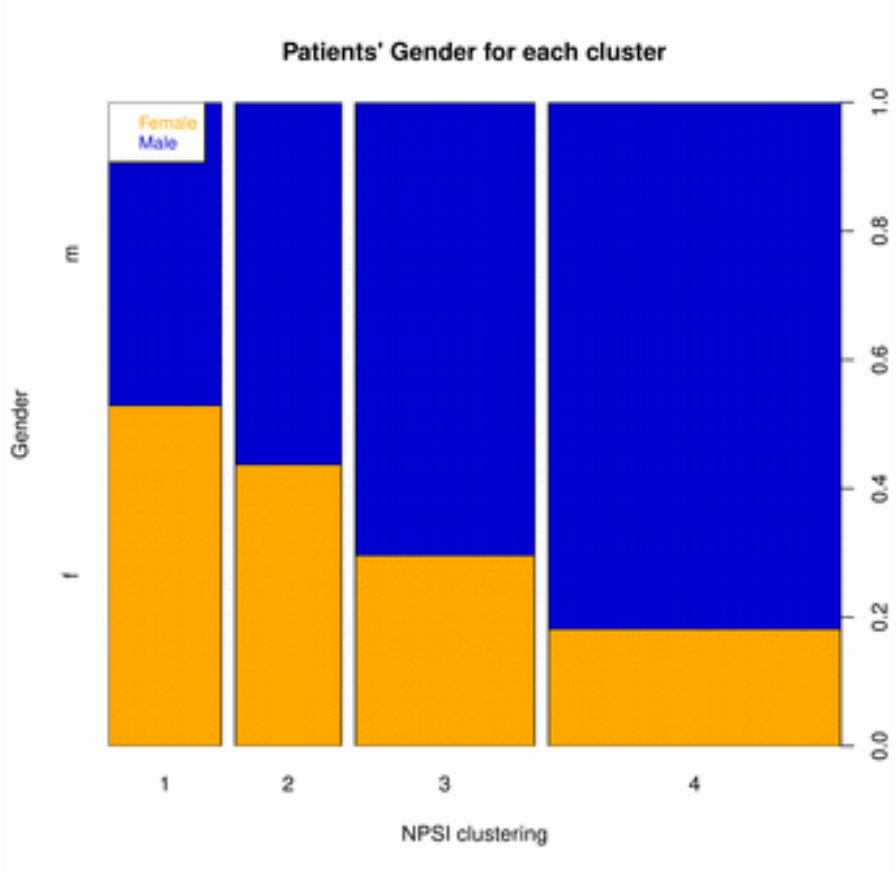


Figure 30: Age and gender are highly associated with cluster assignment.

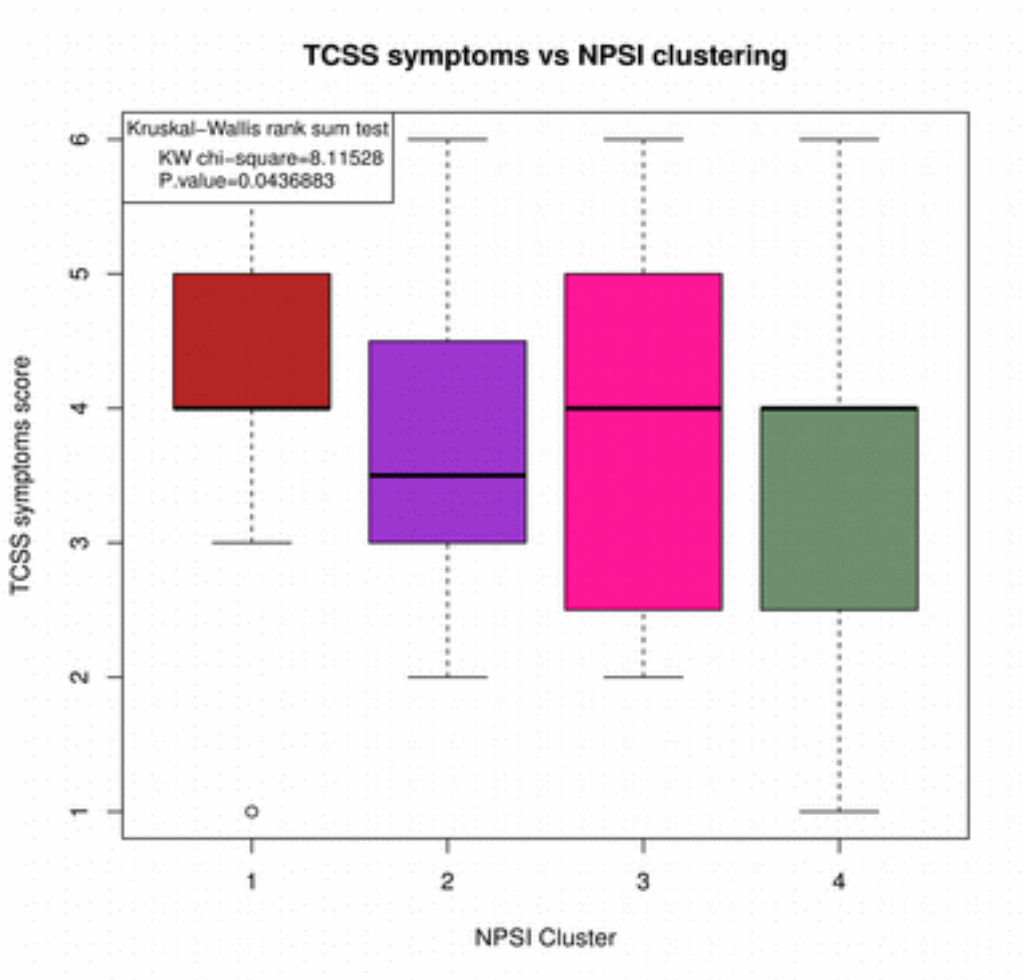


Figure 31:TCSS symptoms sub score has a significant effect on the on grouping of patients according to clusters.

The non-parametric Kruskal-Wallis test showed that patients of different clusters come from populations with significantly differently distributed NPSI subscores, figure 32. Patients clustered in the 3rd cluster had significantly higher Paresthesia/Dysesthesia than patients in cluster 2 (two way students t.test, p.value = 0.0003039) and significantly higher paroxysmal pain (two way students t.test, p.value = 0.01929). On the other hand patients in cluster 2 had slightly higher median values of pain intensity as assessed from the pain diary, NPSI and DN4 and significantly higher median values for Spontaneous and Evoked pain, figure 32.

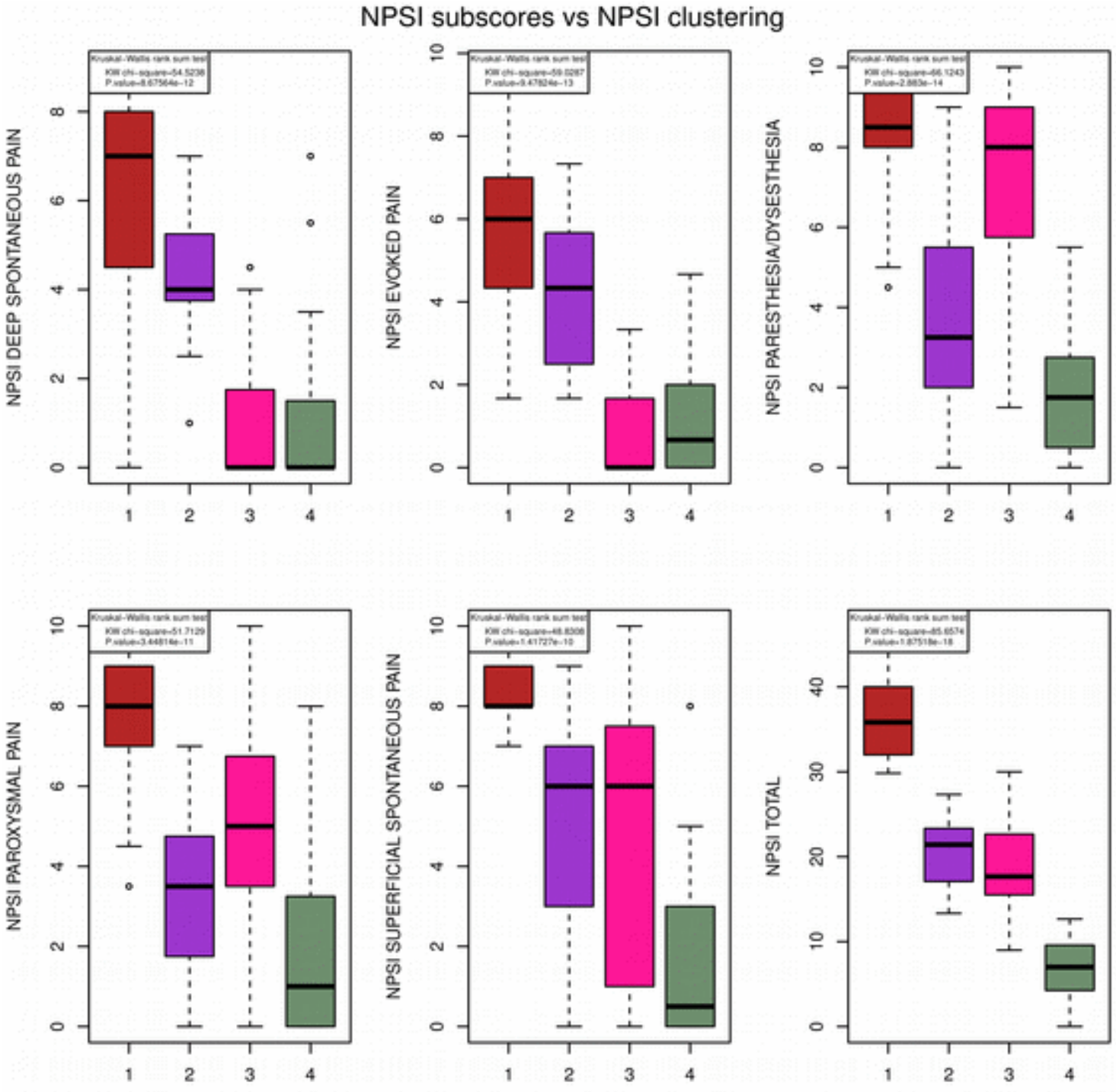


Figure 32: NPSI scores have significant effect on cluster assignment. Moreover clustering revealed distinct modalities of pain obscured from the NPSI total score. Namely cluster 3 has significantly higher paroxysmal pain and paresthesia/ dysesthesia than cluster 2, although cluster 2 has significant higher spontaneous and evoked pain and also higher median values in the total NPSI score.

Conclusion

We analysed data from patients with diabetic neuropathy, a very common aetiology of neuropathy and neuropathic pain. First we confirmed that DN4 questionnaire and the TCSS score and more specifically the TCSS symptoms subscore can distinguish very well between patients with painful neuropathy and painless neuropathy. QST had some mild correlation with these scores but also had significant correlation with loss of intraepidermal nerve fibre density (IENFD). On the other hand tests that were significantly better in screening patients of painful neuropathy had a high correlation with HbA1c blood concentration. Moreover the TCSS symptoms subscore is very highly correlated to the MRC sensory score. Age and sex of the patient was found to be a very important parameter related to pain intensity. Although there was no correlation between age or sex and IENFD, HbA1c or BMI.

In patients with painful neuropathy we found that the NPSI score is highly correlated to the DN4 values, (even though DN4 was designed as a screening test) and the TCSS symptoms subscore. Moreover higher pain intensity assessed by the NPSI average total score was highly correlated to HbA1c and not to IENFD. More interestingly we were able to distinguish 3 or 4, depending on the inclusion of categorical variables assessing paroxysmal pain, distinct somatosensory profiles in the form of varimax rotated principal components, with consistently higher contributions of certain factors of the NPSI questionnaire. Namely superficial pain and paresthesia/dysesthesia; Deep spontaneous pain; Evoked pain and if we included the categorical variables, spontaneous paroxysmal pain. These components were able to separate patients according to pain intensity and they were highly correlated to HbA1c, the age and gender of the patient. In addition, the pain dimension / principal component of superficial pain and paresthesia / dysesthesia had higher negative correlation to age, while the pain dimension / principal component of paroxysmal pain had higher positive correlation to HbA1c.

Moreover we were able to perform unsupervised clustering based on these components which reflected very well the pain intensity. In contrast

QST failed to capture the phenotype of pain intensity and was not able to separate patients having painful neuropathy in any meaningful way related to pain intensity. However, it identified difference between patients of painful and non-painful neuropathy. These distinct clusters based on NPSI data were associated with age, sex, HbA1c and pain intensity. But more interestingly they represented distinct somatosensory profiles of patients having higher intensity in certain modalities of painful neuropathy, figure 32. These profiles are not related with different underlying diseases causing neuropathic pain, but they were rather somatosensory profiles related to pain intensity within only one aetiology of painful neuropathy, diabetic neuropathy. This finding may suggest that symptoms of painful neuropathy emerge in highly diverged forms and combinations and thus require more specialised treatment as well as more research in order to identify whether these clusters are associated with distinct gene expression patterns.

Conclusions and future work

In this thesis we have presented a series of bioinformatics approaches for the better understanding of pain, in both its molecular signature and its divergent phenotypes. Transcriptional profiling technologies have been used for the identification of pain related genes and their expression patterns over the last 15 years. From the first microarray studies (Costigan et al., 2002), the pain genes database (Lacroix-Fralish et al., 2007) and the identification of over-represented GO biological processes of pain (Lötsch et al., 2013) to systems biology of pain (Perkins et al., 2013) and the usage of recent high throughput sequencing technologies (Dawes et al., 2014, 2014). Profiling technologies have identified hundreds of genes which are differentially expressed in pain relevant tissue and involved in pain.

RNA-sequencing (Wang et al., 2009), a very promising transcriptomics technology, has allowed us to investigate and report the relative abundance of thousands of genes between conditions of interest with unprecedented dynamic range and the ability to identify novel genes as pain mediators. Less than 8 years ago a new class of non-protein coding RNA, Long non-coding RNAs started attracting the focus of many researchers studying their function and evolution (Guttman et al., 2009; Ponting et al., 2009). As a consequence bioinformatic pipelines were developed in order to identify novel LncRNAs and study their expression pattern (Ilott and Ponting, 2013). Most of these approaches are based on the breakthrough of modern high throughput sequence by synthesis technologies, usually RNA-sequencing.

Only 3 years ago, antisense LncRNAs were identified as possible mediators of pain (Han and Jan, 2013). A novel LncRNA, expressed in the dorsal root ganglion, antisense of the Kcna2 gene was found to be implicated in neuropathic pain by silencing the expression of the Kcna2 gene (Zhao et al., 2013). Another antisense LncRNA was found to regulate

the expression of the Scn9a pain gene on the opposite strand of the genome (Koenig et al., 2015).

In this study we contributed to the identification of novel LncRNAs that are putative pain mediators. We also study the transcriptional changes involved under pain models of peripheral neuropathy both for known genes and novel LncRNAs.

Our main focus is the transcriptional profiling of neuropathic pain. Thus we presented methods and results from two studies of well established animals models of neuropathic pain and a study dealing with clinical data and self reported pain-questionnaires, which reveals the divergent phenotypes and distinct qualities of painful neuropathy.

Two years ago, a study revealed distinct patterns of symptoms that characterize somatosensory profiles of neuropathic pain based on the NPSI pain questionnaire and Quantitative Sensory Test (Freeman et al., 2014). This study identified four clusters with distinct pain characteristics profiles. Moreover, a large-scale observational study of patients suffering from diabetic neuropathy (Themistocleous et al., 2016) revealed that the severity of neuropathic pain was correlated with higher HbA1c and that diabetic neuropathy is characterised by hypo-sensitivity, which is higher for patients suffering from moderate and severe neuropathic pain. On the other hand paradoxical heat sensation was not a discriminatory feature of painful diabetic neuropathy. We have contributed to the study of distinct somatosensory profiles of diabetic neuropathy which are correlated with clinical parameters and the neuropathic pain severity.

We present the main conclusions and future work below.

Identifying LncRNAs from RNA-seq data

This thesis has been largely focused on developing an efficient strategy for reconstructing gene models of LncRNAs using RNA-seq data in an efficient way, without reconstructing the whole transcriptome, and then

performing a differential expression analysis of these LncRNAs. The goal was to identify a subset of novel LncRNAs, sufficiently and consistently expressed and DE between conditions of interest.

We have provided a customised strategy for predicting gene models of novel LncRNAs without executing the inherently difficult and computationally expensive task of reconstructing the whole transcriptome. Moreover this strategy identifies LncRNAs which are likely to exist and be functionally important as most of them are spliced and multi-exonic and they are sufficiently and consistently expressed and differentially expressed between conditions of interest. The prediction of yet unknown models of LncRNAs remains a very difficult task. Genomic contamination; erroneous base calling in RNA-seq; reads mapping to multiple positions; spurious splicing junctions and an inherent difficulty to identify transcription start and end sites using RNA-seq can generate many false positives and noisy results that can affect all downstream analysis. That is why we decided to apply a series of stringent filters in order to screen genomic loci appearing to have non-canonical transcription that can generate LncRNAs. RNA-sequencing with more than 50 million reads per sample and stranded 100bp paired-end reads was sufficient to perform an *ab initio* identification of the vast majority (about 80%) of expressed LncRNAs found in ENSEMBL/GENCODE annotation. The ones missed were the ones that were so lowly expressed that DE analysis would nevertheless be very difficult to produce confident estimations of LFC. We should note that increased sequencing depth and a better annotated genome, compared to the rat genome, gave much higher dynamic range of transcription strength between highly expressed protein coding genes and lowly expressed predicted LncRNAs.

In **chapter 2: Methods**, we present a computational pipeline which identifies novel LncRNAs in an automated way, using an intuitive strategy of combining a coverage threshold approach of un-annotated genomic regions with de novo identified splicing junctions from a gapped aligner. As it is crucial for downstream analysis, the method produces, as an output, a

set of predicted LncRNAs, in the form of a file following the GTF format which is compatible and ready to be used with state of the art counting techniques and count-based data DE analysis algorithms.

We eventually established a pipeline that transforms raw input of RNA-seq fragments to predicted LncRNAs with good agreement with the ENSEMBL/GENCODE pipeline's prediction of expressed intergenic and antisense LncRNAs in the rat and mouse DRG.

We should note again that predicting novel gene models and/or LncRNAs is not a trivial task. Sequencing noise, random effects, the quality and complexity of genome assemblies and annotations and the inherent inability of RNA-sequencing to accurately determine transcription start and end sites and absolute transcript abundance contributes to the difficulty of the task. Immediate future work will include the addition to the pipeline of other sources of information like cap analysis gene expression (CAGE) data (Okazaki et al., 2002) for determining transcription start sites and ChIP-seq data to determine chromatin modifications related to actual transcription. These would allow for more confident predictions.

Transcriptional profiling of rodents DRG

In **chapter 3: Transcriptional changes of protein coding genes and novel LncRNAs in rat's DRG**, we investigated transcriptional changes in rat DRG for the SNT pain model. We identified the contribution of ion channels, agents of neuron regeneration and development and opioid receptors in this well induced pain state.

We also predicted hundreds of novel LncRNAs expressed in rat DRG. 21 of these LncRNAs were significantly DE antisense of significantly DE pain genes. One LncRNA, antisense of the Kcnj9 voltage gated potassium channel, has opposite expression pattern with the protein coding on the opposite strand. Thus we hypothesize that it might regulate its expression and thus contribute to neuropathic pain.

We also identified 7 intergenic LncRNAs significantly DE and highly correlated with their adjacent pain gene. We hypothesize that these

LncRNAs may contribute to neuropathic pain by regulating their neighbouring pain gene *in cis*.

In **chapter 4: Transcriptional changes in DRG of two mouse strains experiencing high and low induced hypersensitivity** for the SNI pain model, a follow up study of two mouse strains, BALB/c – high pain and B10.D2 – low pain, after the SNI pain model we found that the high pain strain has more significantly DE genes than the low pain strain and that the high pain strain is much more similar to rat in terms of transcriptional changes for the pain model.

Moreover, the high pain strain has more prominent dysregulation of voltage gated potassium and sodium channels. Functional enrichment analysis revealed that biological processes of neuron regeneration and development, axon guidance, regulation of ion channels, signalling, response to stimuli, learning and memory were highly enriched. Thus the common core of enriched biological processes in both strains is almost identical with the pain wheel of enriched biological processes for pain genes (See Introduction , section Pain at the molecular level) (Lötsch et al., 2013).

Moreover genes related to axon guidance, neuron development and regeneration, potassium ion transport, chemokine regulation and immune system had significantly different responses for the pain model between strains. We hypothesize that these processes contribute to the significant difference in induced allodynia for the SNI model between the two strains.

Regarding predicted LncRNAs, we again found that more of these were DE in the high pain strain than in the low pain strain. 12 LncRNAs antisense of pain genes were syntenically conserved between mouse and rat. 4 intergenic LincRNAs with a pain gene as their closest genomic neighbour were syntenically conserved between mouse and rat. We found in both species the known antisense LncRNAs on the opposite strand of Scn9a and Kcna2 genes.

Two pairs of protein coding genes Nalcn and Tshz – and their respective antisense LncRNA were significantly DE only in the high pain

strain. Thus we hypothesize that these antisense LncRNAs might be related with the significant differences in the intensity of pain observed between the mouse strains. Moreover we identified a significantly correlated and significantly DE LincRNA with its closest neighbour, the Oprd1 opiod receptor in mouse and with the Oprm1 opioid receptor in rat. These two receptors interact and form hetero-dimers. Thus we hypothesize that these LincRNAs might contribute to neuropathic pain by regulating these opioid receptors in both species.

Most of the LncRNAs predicted in both species were bi-exonic. About half of the LncRNAs predicted were antisense of annotated protein coding genes. In addition most of the DE LincRNAs were in close proximity to protein coding genes. There were no significantly DE LincRNAs very far away from known genes. Most of the LncRNAs predicted showed low coding potential. However we also predicted gene models which showed positive coding potential. Some of these were DE. These might be processed pseudogenes or not yet annotated protein coding genes.

These studies gave insights regarding transcriptional changes both in protein coding genes and LncRNAs in rodents' DRG for well established pain models. Immediate future work is to validate these predictions using Q-PCR and infer the underlying biology of their hypothesized function. We are preparing a manuscript presenting these results and upon publication these will be uploaded to Pain Networks in order to contribute to the publicly available transcriptomics data and the co-expression networks for pain.

Divergent phenotypes of painful neuropathy

In **chapter 5: Clustering of patients with diabetic neuropathy reveals distinct neuropathic pain dimensions**, we changed our scope in studying neuropathic pain and presented results from an analysis of data from human patients with diabetic neuropathy. We analysed both clinical data and data from quality of life – pain questionnaires. First we confirmed

that DN4 and TCSS are very effective in screening neuropathic pain but are also correlated with neuropathic pain intensity as measured by NPSI. On the other hand QST cannot capture the neuropathic pain intensity phenotype.

More severe pain was highly associated with higher concentrations of HbA1c, younger age and female gender. All these variables were also highly associated with NPSI, DN4 and TCSS symptom scores. On the other hand QST was correlated with IENFD but neither QST scores nor nerve fibre density (IENFD) were associated with pain intensity.

Moreover, clustering of patients based on the NPSI data revealed 4 distinct somatosensory profiles associated with pain intensity. Higher scores in deep spontaneous pain and evoked pain were characteristic of the two clusters of patients with more intense neuropathic pain. On the other hand high scores of paresthesia / dysesthesia and paroxysmal pain were characteristics of the cluster of moderate pain. Finally the cluster of mild pain had low scores in all NPSI sub-scores. These clusters were associated with the 7-day pain diary scores, TCSS symptom scores, the gender, the age and HbA1c concentration.

Immediate future work will include sequencing of the skin biopsy patches used in this study in order to establish possible transcriptional differences between the above clusters.

In this thesis we presented a series of bioinformatic attempts aiming to the better understanding of pain. Technological advances and optimisation of current protocols will allow for better usage of the vast amounts of data available. RNA-seq has been established as the standard transcriptional profiling tool and standard analytical workflows are becoming readily available, more accurate and more accessible. The computational resources for identifying LncRNAs are slowly entering a more mature phase and their function can be further investigated, although serious biological and computational challenges remain ahead. New technologies like single cell sequencing (Gawad et al., 2016) can perform targeted profiling of specific cells in complex systems like pain and

nociception. This can further de-convolute the contribution of different biological systems in pain phenotype.

Of course all this data will eventually bring forward new challenges of bioinformatics and computational biology. We would only be able to overcome these challenges by establishing broad collaborations involving a diverse spectrum of scientists, studying both animal models of pain and humans.

Appendix 1

PC1	PC1_symbol	PC2	PC2_symbol
ENSRNOG00000020557	Ryr1	ENSRNOG00000007304	Herc3
ENSRNOG00000010079	Car3	ENSRNOG00000028992	Acan
ENSRNOG00000052355	NA	ENSRNOG00000009322	Ccdc126
ENSRNOG00000017645	Mylpf	ENSRNOG00000009170	Dmxl2
ENSRNOG00000051895	NA	ENSRNOG00000026447	Ptchd2
ENSRNOG00000013262	Myl1	ENSRNOG00000018827	Htr7
ENSRNOG00000059651	NA	ENSRNOG00000020650	Slc17a7
ENSRNOG00000061829	NA	ENSRNOG00000014323	Extl2
ENSRNOG00000022637	NA	ENSRNOG00000006037	NA
ENSRNOG00000019627	Mybpc2	ENSRNOG00000007999	Abra
ENSRNOG00000004583	Mb	ENSRNOG00000014302	Dlgap3
ENSRNOG00000020276	Tnni2	ENSRNOG00000032472	Adgrg2
ENSRNOG00000016837	Ckm	ENSRNOG00000019318	Syt3
ENSRNOG00000019745	Actn3	ENSRNOG00000009465	Sfrp2
ENSRNOG00000015155	Tnnc2	ENSRNOG00000012015	Lrrc49
ENSRNOG00000047124	NA	ENSRNOG00000003756	Lancl3
ENSRNOG00000020332	Tnnt3	ENSRNOG00000020164	NA
ENSRNOG00000005154	Fam150b	ENSRNOG00000042477	Clrn1
ENSRNOG00000056493	Mybpc1	ENSRNOG00000026577	Cpne4
ENSRNOG00000057701	Myom1	ENSRNOG00000014453	Anxa5
ENSRNOG00000058083	NA	ENSRNOG00000017539	Mmp9
ENSRNOG00000031782	Col6a4	ENSRNOG00000018054	F2rl2
ENSRNOG00000010803	Gabra5	ENSRNOG00000037695	Sgpp2
ENSRNOG00000010478	Serpina3n	ENSRNOG00000009577	Ndst4
ENSRNOG00000009907	Mmp8	ENSRNOG00000021015	Sbsn
ENSRNOG00000009465	Sfrp2	ENSRNOG00000014840	Gna14
ENSRNOG00000008478	Mmp13	ENSRNOG00000012038	Htr1d
ENSRNOG00000028992	Acan	ENSRNOG00000054795	NA
ENSRNOG00000040350	Mir675	ENSRNOG00000052514	NA
ENSRNOG00000061739	Klrk1	ENSRNOG00000025463	LOC100125362
ENSRNOG00000007178	Cd8a	ENSRNOG00000038004	Zfp804a
ENSRNOG00000006926	Atp6v0d2	ENSRNOG00000046468	Ptgfr
ENSRNOG00000015411	Apobec1	ENSRNOG00000057501	Fam81a
ENSRNOG00000014610	Anpep	ENSRNOG00000055401	Kcnc1
ENSRNOG00000018286	Chrna1	ENSRNOG00000038297	Plekhd1
ENSRNOG00000021029	Hamp	ENSRNOG00000006617	NA
ENSRNOG00000021062	Fxyd5	ENSRNOG00000012196	Asah2
ENSRNOG00000004641	Sstr4	ENSRNOG00000011146	Gyg1
ENSRNOG00000020845	Tyrobp	ENSRNOG00000061304	Atp2b3
ENSRNOG00000015076	Cyp26b1	ENSRNOG00000030238	Fndc5
ENSRNOG00000037331	Cd33	ENSRNOG00000011285	Zdhhc22
ENSRNOG00000033564	Cfd	ENSRNOG00000011550	Kcnab2
ENSRNOG00000007918	Tbxas1	ENSRNOG00000019648	NA
ENSRNOG00000016037	Mafb	ENSRNOG00000042788	NA
ENSRNOG00000015562	Cdh17	ENSRNOG00000029478	Cyp4f39
ENSRNOG00000024082	Gldn	ENSRNOG00000007727	Lhfpl4
ENSRNOG00000025691	Pla2g7	ENSRNOG00000058938	Camkv
ENSRNOG00000008015	Fos	ENSRNOG00000050419	Avil
ENSRNOG00000008409	Myo1f	ENSRNOG00000006639	Scn9a
ENSRNOG00000004578	Cthrc1	ENSRNOG00000018191	Oprm1

Table 1: Top 50 ENSEMBL genes contributing to PC1 and PC2 in rat DRG

Appendix 2

LncRNA	ENSEMBL ID	Gene symbol	lnc_lfc	lnc_pvalue	gene_lfc	gene_pvalue	cpc
chr1:37747286-37751337(-)	ENSRNOG00000017601	Srd5a1	-0.0447807219	0.9190998397	0.4351760184	0.0366284775	-1.24251
chr1:81878371-81879779(+)	ENSRNOG00000018454	Apoe	-0.566801053	0.0567754688	-0.3457943355	0.0702242035	-1.1206
chr1:123544997-124296791(-)	ENSRNOG00000010146	Ndn	-0.6667793509	0.0606260842	-0.0455532867	0.8217523885	-1.1682
chr1:177247980-177271718(-)	ENSRNOG00000017679	Cckbr	2.112658187	1.44658406879268E-06	1.9467315106	3.48587574213042E-07	-0.88056
chr1:191162390-191236103(+)	ENSRNOG00000011130	Calca	-1.820836666	0.0039226154	-0.8165221175	0.0975339719	-0.538208
chr1:222437062-222437743(+)	ENSRNOG00000020206	Ctsd	-0.6235290917	0.304732215	-0.3633521581	0.0988581346	-1.15509
chr1:229209248-229210864(+)	ENSRNOG00000021150	Plcb3	0.485775811	0.522226726	-0.3516802804	0.0977593436	-1.12512
chr1:245206514-245209217(+)	ENSRNOG00000017469	Anxa1	1.2247686954	0.0628253085	0.4531432481	0.0990798605	-1.00017
chr10:40573511-40574610(+)	ENSRNOG00000012840	Sparc	0.693080439	0.0309146894	0.5219163833	0.0004593778	-0.972695
chr10:49346229-49346991(-)	ENSRNOG00000003338	Pmp22	-0.1177015925	0.7477371661	0.236004643	0.3916407903	-1.19616
chr10:56591574-56650214(+)	ENSRNOG00000027037	Alox12	0.1314851106	0.563538108	0.2755623591	0.4643322887	-0.0382499
chr10:56761324-56812677(-)	ENSRNOG00000019308	Arrb2	-0.7070928775	0.0629842467	-0.3152262668	0.0469287398	-1.11061
chr10:62857461-62860411(+)	ENSRNOG00000003476	Slc6a4	-0.6717379382	0.2842524415	NA	NA	-0.372492
chr10:81937955-81950750(+)	ENSRNOG00000002948	Abcc3	0.6430729974	0.1193131905	0.4573801324	0.0104184609	-0.800468
chr10:81950724-81953548(+)	ENSRNOG00000002981	NA	0.1272952672	0.5374914338	-0.0268674599	0.9424883437	-0.575908
chr10:90227150-90682919(-)	ENSRNOG00000002753	Adam11	0.073446823	0.6725234564	-1.013439415	0.0001701342	-1.15469
chr10:96310354-96310917(+)	ENSRNOG00000003491	Prkca	-0.1990222363	0.7690545773	-0.5076312503	0.0053541231	-1.23327
chr11:29078952-29091061(+)	ENSRNOG00000001606	NA	-0.392644359	0.0227719465	-0.1424088152	0.2832251035	-0.886432
chr11:31735382-31864076(+)	ENSRNOG00000001575	Grik1	-1.2187665724	0.0008949842	-1.4425118869	0.0026368395	-0.940265
chr11:36455832-36456560(+)	ENSRNOG00000001704	Runx1	1.0564240662	0.0674304349	0.1714696229	0.3626048906	-0.768232
chr11:62764709-62826162(-)	ENSRNOG00000001528	Gap43	-0.202150834	0.4350106104	1.1099123845	7.66981630138424E-05	-0.100841
chr11:71932813-71946183(+)	ENSRNOG00000002229	Adcy5	-0.3028927596	0.3669153667	0.0050502205	0.9787693274	-1.01178
chr12:1148261-1148984(+)	ENSRNOG00000001090	Stard13	0.9088934954	0.0425706885	1.0436792285	0.0002948938	-1.18939
chr12:41201037-41219792(+)	ENSRNOG00000001300	P2rx4	0.824750745	0.2159980941	0.259414065	0.3010495352	-0.812486
chr13:52205634-52281411(+)	ENSRNOG00000003927	Cd55	-0.7126622392	0.2030060805	-1.2274586192	0.0093133909	0.573113
chr13:95223994-95228919(+)	ENSRNOG00000007645	Kcnj9	1.1049036718	6.56002103938819E-05	-1.2226296115	0.0003202419	0.459104
chr14:8191831-8193162(-)	ENSRNOG00000002079	Mapk10	0.2401930381	0.7403921646	-0.8442717744	0.0004100803	-1.50245
chr14:34155103-34162352(-)	ENSRNOG00000002164	Nmu	-0.4489053786	0.1308590754	NA	NA	-1.06048
chr14:42928354-42928641(+)	ENSRNOG00000002343	Uchl1	-0.4731225392	0.1839489513	-0.3091734109	0.1359027758	-1.05083
chr14:71785314-71786402(+)	ENSRNOG00000003069	Cd38	-0.6344039521	0.3367613316	0.2377024038	0.5153810696	-1.09667
chr15:91530679-91541466(+)	ENSRNOG00000010997	Ednrb	-0.0703856724	0.8215235733	-0.0640513458	0.7906606809	-1.1472
chr15:107356568-107361999(+)	ENSRNOG00000010064	Abcc4	-0.9388852963	0.0034811028	0.0287800233	0.8511005855	-1.02241
chr16:11585122-11585914(+)	ENSRNOG00000020155	Mapk8	0.572410107	0.4302871179	-0.2941889671	0.0097183305	-1.01858
chr16:19636469-19640494(+)	ENSRNOG00000016892	Nr2f6	0.2024489197	0.5545590293	0.3143276744	0.0190656913	-1.10719
chr16:37462470-37491708(+)	ENSRNOG00000049792	Glra3	0.2838969898	0.7061792884	NA	NA	-1.5332
chr19:11642610-11644968(+)	ENSRNOG00000019482	Gnao1	-0.3340469383	0.5796951191	-0.6652059467	0.0078647986	-0.892065
chr2:41590850-41591127(+)	ENSRNOG00000012471	Thbs4	0.9987025993	0.1449421069	0.8050443217	0.0208455771	-1.1965
chr2:63334966-63356823(-)	ENSRNOG00000013963	Il6st	0.2337646709	0.7210196331	0.3063017582	0.0126989676	-1.15433
chr2:170727905-170730261(-)	ENSRNOG00000014232	P2ry1	-0.1317282099	0.7404354621	-0.3621258376	0.1032307581	-0.738574
chr2:199111645-199115704(+)	ENSRNOG00000028589	Gria2	-1.5735337199	8.73053258903914E-05	-1.7163376183	8.44767579830593E-08	-0.97916
chr2:223469266-223471786(+)	ENSRNOG00000030019	Atp1a1	-1.5871758913	3.48442041275082E-05	-1.3611462919	0.000031922	-1.14917
chr20:13639959-13659017(-)	ENSRNOG00000001216	Trpm2	0.0218240816	0.9655273051	0.6565867731	1.99392721612971E-05	-0.739585
chr20:45787077-45787666(-)	ENSRNOG00000000599	Lama4	0.6684365109	0.2970693435	0.7159466344	0.001954655	0.378354
chr20:46160790-46161175(-)	ENSRNOG00000000596	Fyn	0.4435010339	0.5537561776	0.4118310479	0.0015142443	-1.01299
chr3:59057403-59276093(+)	ENSRNOG00000006639	Scn9a	-2.357018731	1.42524585316199E-09	-1.0866688096	1.44524593696598E-05	-0.815863
chr3:69041756-69043454(+)	ENSRNOG00000001548	Nfe2l2	0.8870856496	0.1233891464	0.1634197457	0.2048763399	-0.232069
chr3:127260488-127283308(-)	ENSRNOG00000014152	Kcnip3	-1.5362755276	0.00000623	-1.6266979042	4.14361222986572E-05	-1.15812
chr4:7251609-7306887(+)	ENSRNOG00000008380	Asic3	-0.296239585	0.3378583875	-0.2285426371	0.1836174061	0.353415
chr4:9636853-9646624(-)	ENSRNOG00000021441	Reln	1.2405189319	0.0029984059	0.7327351616	0.0018675684	-1.30911
chr4:162569914-162651062(+)	ENSRNOG00000005615	Gadd45a	0.9741379114	0.0011496768	1.0547848024	0.0022037369	-0.434463
chr4:224270790-224274242(+)	ENSRNOG00000014294	Ptpn6	-0.1183917742	0.8562262094	0.8228388998	0.003274947	-1.54357
chr4:224746318-224747269(-)	ENSRNOG00000019219	Vamp1	-0.6141431576	0.2040033788	-1.5285410153	0.0006560966	-0.744079
chr5:3719233-3765605(-)	ENSRNOG00000007354	Trpa1	-1.0607031702	0.0005619469	-0.8457661675	0.0203168171	-0.814652
chr5:102448464-102448826(-)	ENSRNOG00000029318	Tyrp1	-0.1928496338	0.7902991047	-0.9854320236	0.0236335401	-1.27712
chr5:154402173-154493191(-)	ENSRNOG00000013231	Ptafr	0.677320791	0.0271052464	0.6466799012	0.0313546671	-1.52982
chr6:9546755-9553254(-)	ENSRNOG00000015603	NA	0.020837421	0.969466824	-0.9300543601	2.49781655973716E-05	-1.11413
chr7:126391081-126394210(-)	ENSRNOG00000021463	Ppara	-0.0300122855	0.900186327	0.1982840405	0.4581352938	0.680499
chr7:140661997-140671754(-)	ENSRNOG00000007346	Grasp	0.3462366023	0.3027295777	0.3944661433	0.1206041916	-1.20061
chr8:2515108-2772995(-)	ENSRNOG00000007372	Casp1	0.0935904095	0.8743411701	0.7067930217	0.0098913233	-1.89693
chr8:32323519-32324313(+)	ENSRNOG00000047179	Aplp2	-1.6363414407	0.0115691932	-0.7368788555	4.68583620320112E-05	-1.12105
chr8:48034756-48064293(-)	ENSRNOG00000016221	Scn2b	-0.8096772721	0.0208522022	-0.0207163123	0.9410233127	-0.925869
chr8:57331729-57332065(+)	ENSRNOG00000000196	Cyp19a1	0.2466031844	0.7251451474	NA	NA	-1.29188
chr8:88667082-88711088(+)	ENSRNOG00000013042	Htr1b	1.1366503299	5.71176513352015E-05	0.3209479749	0.0322417882	-1.06466
chr8:103496790-103720412(+)	ENSRNOG00000011501	Atp1b3	0.0227119781	0.9141564307	-0.7630008751	0.0001102393	-1.72487

Table 1: All LncRNAs in rat DRG antisense of pain genes

Appendix 3

	PC1	PC1_symbol	PC2	PC2_symbol
1	ENSMUSG00000036395	Glb1l2	ENSMUSG00000023046	Igfbp6
2	ENSMUSG00000110631	NA	ENSMUSG00000049134	Nrap
3	ENSMUSG00000024366	Gfra3	ENSMUSG00000017344	Vtn
4	ENSMUSG00000001435	Col18a1	ENSMUSG00000030108	Slc6a13
5	ENSMUSG00000041607	Mbp	ENSMUSG00000064387	Snora73a
6	ENSMUSG00000004098	Col5a3	ENSMUSG00000075307	Klhl41
7	ENSMUSG00000045573	Penk	ENSMUSG00000019577	Pdk4
8	ENSMUSG00000027273	Snap25	ENSMUSG00000078234	Klhdc7a
9	ENSMUSG00000040152	Thbs1	ENSMUSG00000111340	NA
10	ENSMUSG00000092805	NA	ENSMUSG00000008845	Cd163
11	ENSMUSG00000049176	Frmpd4	ENSMUSG00000010122	Slc47a1
12	ENSMUSG00000045589	Frrs1l	ENSMUSG00000045039	Megf8
13	ENSMUSG00000026042	Col5a2	ENSMUSG00000024049	Myom1
14	ENSMUSG00000021219	Rgs6	ENSMUSG00000041559	Fmod
15	ENSMUSG00000000223	Drp2	ENSMUSG00000030107	Usp18
16	ENSMUSG00000030108	Slc6a13	ENSMUSG00000026904	Slc4a10
17	ENSMUSG00000042286	Stab1	ENSMUSG00000024529	Lox
18	ENSMUSG00000064945	Rny3	ENSMUSG00000044938	Klhl31
19	ENSMUSG00000095079	NA	ENSMUSG00000032589	Bsn
20	ENSMUSG00000056306	Sertm1	ENSMUSG00000031722	Hp
21	ENSMUSG00000074899	NA	ENSMUSG00000027737	Slc7a11
22	ENSMUSG00000040690	Col16a1	ENSMUSG00000040055	Gjb6
23	ENSMUSG00000020396	Nefh	ENSMUSG00000031765	Mt1
24	ENSMUSG00000031538	Plat	ENSMUSG00000045573	Penk
25	ENSMUSG00000033595	Lgi3	ENSMUSG00000025153	Fasn
26	ENSMUSG00000025348	Itga7	ENSMUSG00000056328	Myh1
27	ENSMUSG00000076609	NA	ENSMUSG00000001348	Acp5
28	ENSMUSG00000024621	Csf1r	ENSMUSG00000038541	Srd5a2
29	ENSMUSG00000064337	NA	ENSMUSG00000036814	Slc6a20a
30	ENSMUSG00000015647	Lama5	ENSMUSG00000023993	Treml1
31	ENSMUSG00000044071	Fam19a2	ENSMUSG00000004891	Nes
32	ENSMUSG00000003746	Man1a	ENSMUSG00000056174	Col8a2
33	ENSMUSG00000036594	H2-Aa	ENSMUSG00000031762	Mt2
34	ENSMUSG00000035202	Lars2	ENSMUSG00000050578	Mmp13
35	ENSMUSG00000041020	Map7d2	ENSMUSG00000034664	Itga2b
36	ENSMUSG00000013584	Aldh1a2	ENSMUSG00000070385	Ampd1
37	ENSMUSG00000090667	Gm765	ENSMUSG00000021390	Ogn
38	ENSMUSG00000073418	C4b	ENSMUSG00000035202	Lars2
39	ENSMUSG00000047344	Lancl3	ENSMUSG00000037736	Limch1
40	ENSMUSG00000047216	Cdh19	ENSMUSG00000026051	1500015O10Rik
41	ENSMUSG00000033066	Gas7	ENSMUSG00000032648	Pygm
42	ENSMUSG00000024164	C3	ENSMUSG00000029843	Slc13a4
43	ENSMUSG00000063011	Msln	ENSMUSG00000024471	Myot
44	ENSMUSG00000003477	Inmt	ENSMUSG00000026697	Myoc
45	ENSMUSG00000026712	Mrc1	ENSMUSG00000016255	Tubb1
46	ENSMUSG00000018217	Pmp22	ENSMUSG00000030116	Mfap5
47	ENSMUSG00000045672	Col27a1	ENSMUSG00000064945	Rny3
48	ENSMUSG00000030218	Mgp	ENSMUSG00000013584	Aldh1a2
49	ENSMUSG00000032854	Ugt8a	ENSMUSG00000029373	Pf4
50	ENSMUSG00000046157	Tmem229b	ENSMUSG00000098178	NA

Table 1: Top 50 genes contributing to PC1 and PC2 for the BALB/c mouse strain

	PC1	PC1_symbol	PC2	PC2_symbol
1	ENSMUSG00000032332	Col12a1	ENSMUSG00000061762	Tac1
2	ENSMUSG00000083161	NA	ENSMUSG00000024222	Fkbp5
3	ENSMUSG00000109908	NA	ENSMUSG00000032496	Ltf
4	ENSMUSG00000094546	NA	ENSMUSG00000037868	Egr2
5	ENSMUSG00000086825	NA	ENSMUSG00000093843	NA
6	ENSMUSG00000095589	NA	ENSMUSG00000026395	Ptprc
7	ENSMUSG00000076940	NA	ENSMUSG00000035200	Chrnb4
8	ENSMUSG00000026904	Slc4a10	ENSMUSG00000020849	Ywhae
9	ENSMUSG00000076258	NA	ENSMUSG00000032303	Chrna3
10	ENSMUSG00000076672	NA	ENSMUSG00000005952	Trpv1
11	ENSMUSG00000019874	Fabp7	ENSMUSG00000051855	Mest
12	ENSMUSG00000037868	Egr2	ENSMUSG00000010154	Spire2
13	ENSMUSG00000019960	Dusp6	ENSMUSG00000083161	NA
14	ENSMUSG00000042644	Itpr3	ENSMUSG00000026904	Slc4a10
15	ENSMUSG00000076617	NA	ENSMUSG00000030157	Clec2d
16	ENSMUSG00000059824	Dbp	ENSMUSG00000020599	Rgs9
17	ENSMUSG00000093861	NA	ENSMUSG00000042942	Greb1l
18	ENSMUSG00000017344	Vtn	ENSMUSG00000031667	Aktip
19	ENSMUSG00000100510	LOC102636514	ENSMUSG00000026185	Igfbp5
20	ENSMUSG00000025479	Cyp2e1	ENSMUSG00000069919	Hba-a1
21	ENSMUSG00000094491	NA	ENSMUSG00000095130	NA
22	ENSMUSG00000004105	Angptl2	ENSMUSG00000036699	Zcchc12
23	ENSMUSG00000106106	NA	ENSMUSG00000094546	NA
24	ENSMUSG00000076563	NA	ENSMUSG00000070570	Slc17a7
25	ENSMUSG00000064367	ND5	ENSMUSG00000021647	Cartpt
26	ENSMUSG00000066687	Zbtb16	ENSMUSG00000020483	Dynll2
27	ENSMUSG00000078234	Klhdc7a	ENSMUSG00000022054	Nefm
28	ENSMUSG00000032589	Bsn	ENSMUSG00000030790	Adm
29	ENSMUSG00000076677	NA	ENSMUSG00000061535	C1qtnf7
30	ENSMUSG00000027737	Slc7a11	ENSMUSG00000076652	NA
31	ENSMUSG00000038872	Zfhx3	ENSMUSG00000022123	Scel
32	ENSMUSG00000061762	Tac1	ENSMUSG00000056054	S100a8
33	ENSMUSG00000030730	Atp2a1	ENSMUSG00000050963	Kcns2
34	ENSMUSG00000067786	Nnat	ENSMUSG00000006411	Pvrl4
35	ENSMUSG00000053279	Aldh1a1	ENSMUSG00000007682	Dio2
36	ENSMUSG00000035000	Dpp4	ENSMUSG00000019874	Fabp7
37	ENSMUSG00000095889	NA	ENSMUSG00000025270	Alas2
38	ENSMUSG00000032657	Fam189b	ENSMUSG00000046480	Scn4b
39	ENSMUSG00000039488	Cntn5	ENSMUSG00000056071	S100a9
40	ENSMUSG00000051747	Ttn	ENSMUSG00000087382	NA
41	ENSMUSG00000056215	Lrguk	ENSMUSG00000096349	NA
42	ENSMUSG00000020483	Dynll2	ENSMUSG00000094075	NA
43	ENSMUSG00000094433	NA	ENSMUSG00000041556	Fbxo2
44	ENSMUSG00000038193	Hand2	ENSMUSG00000073940	Hbb-b1
45	ENSMUSG00000102364	NA	ENSMUSG00000028369	Svep1
46	ENSMUSG00000044734	Serpinb1a	ENSMUSG00000020701	Tmem132e
47	ENSMUSG00000002944	Cd36	ENSMUSG00000020053	Igf1
48	ENSMUSG00000024650	Slc22a6	ENSMUSG00000096833	NA
49	ENSMUSG00000091345	Col6a5	ENSMUSG00000038319	Kcnh2
50	ENSMUSG00000010122	Slc47a1	ENSMUSG00000067149	Jchain

Table 2: Top 50 genes contributing to PC1 and PC2 for the B10.D2 strain

Appendix 4

LncRNA Genomic Coordinates	ENSEMBL ID	symbol	lnc_lfc	lnc_pvalue	gene_lfc	gene_pvalue	cpc
chr1:131263308-131266930(+)	ENSMUSG00000042349	Ikbke	-0.98520025	0.6592695429	0.3261541797	0.2142890364	-1.16173
chr10:39555935-39560136(-)	ENSMUSG00000019843	Fyn	-0.277984631	0.3544766412	0.1165577866	0.1207782839	-0.472081
chr11:7213635-7261406(+)	ENSMUSG00000020427	Igfbp3	0.3107693246	0.7052722355	0.1947102662	0.5265263297	-1.00746
chr11:55394494-55395613(+)	ENSMUSG00000018593	Sparc	0.154481082	0.9317813792	0.3846903375	0.0044242494	-1.05264
chr11:63150892-63151309(-)	ENSMUSG00000018217	Pmp22	0.2368154814	0.9104057371	0.0367805138	0.9290005349	-0.886487
chr11:66896676-66931091(-)	ENSMUSG00000048070	Pirt	-0.201672067	0.772517491	-0.155669631	0.1423672603	0.907072
chr11:70239888-70246096(+)	ENSMUSG00000000320	Alox12	0.1813909634	0.9004524871	0.4654956959	0.5374670032	0.125631
chr11:73296005-73300736(-)	ENSMUSG00000043029	Trpv3	0.3789321054	0.4348536675	0.052672489	0.9293436747	-0.979906
chr11:81966759-81995840(+)	ENSMUSG00000020704	Asic2	-0.343159024	0.2996970344	-0.091843141	0.6553707748	-0.833933
chr11:83526976-83532680(+)	ENSMUSG00000035042	Ccl5	1.0110523243	0.6322415666	0.7629876473	0.4298909755	-0.685233
chr11:102739325-102762503(-)	ENSMUSG00000020926	Adam11	-0.176744019	0.7599021089	-0.157396544	0.364093453	0.539235
chr11:107935614-107937468(+)	ENSMUSG00000050965	Prkca	0.5167625641	0.8478472381	-0.178048418	0.0338701243	-0.893234
chr13:112505712-112508381(-)	ENSMUSG00000021756	Il6st	0.1297205562	0.9465674986	0.0612605241	0.4975361924	-1.01211
chr14:74638797-74642453(-)	ENSMUSG00000034997	Htr2a	0.4123544967	0.9076411811	0.0251968876	0.9185442832	-1.24073
chr14:103778950-103851424(+)	ENSMUSG00000022122	Ednrb	0.4948442126	0.1528368859	0.145793644	0.2528297697	0.298243
chr15:78081606-78108287(+)	ENSMUSG00000019146	Cacng2	-0.296514822	0.7374523878	-0.57794508	0.0170050333	-0.615593
chr15:78919895-78928253(-)	ENSMUSG00000068220	Lgals1	0.1777692974	0.8991514195	0.6709975096	3.976506E-18	-0.956606
chr15:79071107-79242303(-)	ENSMUSG00000068206	Pick1	0.2698762926	0.785038705	-0.065600513	0.6456907014	0.265443
chr15:101214611-101225267(-)	ENSMUSG00000000531	Grasp	0.098875379	0.9798524065	0.1195690448	0.6817437262	-0.116895
chr15:102138265-102205121(-)	ENSMUSG00000023046	Igfbp6	-0.067812584	0.8698159699	-0.067706189	0.8999149811	0.527199
chr16:85898584-85905300(+)	ENSMUSG00000022894	Adamts5	0.3483300083	0.3330813848	0.000848238	0.997377315	-0.949085
chr16:87934064-87936434(+)	ENSMUSG00000022935	Grik1	-0.921835283	0.6451204509	-0.17453844	0.1805609643	-1.01687
chr16:92690953-92693516(+)	ENSMUSG00000022952	Runx1	0.7466460941	0.5433316449	-0.150059313	0.3051862594	-1.09998
chr16:94752627-94753306(+)	ENSMUSG00000043301	Kcnj6	0.8951291087	0.7257449949	0.0464183082	0.9528614695	-1.04139
chr17:86375800-86379306(-)	ENSMUSG00000045038	Prkce	-0.996940145	0.2785528776	-0.145539461	0.2439022269	-0.899925
chr18:4352995-4368040(+)	ENSMUSG00000024235	Map3k8	-0.54220561	0.7196399978	0.0772844548	0.8491287909	-1.1489
chr18:82405614-82406458(+)	ENSMUSG00000024553	Galr1	0.1235252513	0.9545689021	-0.034555841	0.9451380866	-0.350143
chr19:6969343-6970896(+)	ENSMUSG00000024960	Plcb3	-0.166144246	0.9646489653	-0.161539318	0.3413281225	-0.934247
chr19:22435556-22448608(-)	ENSMUSG00000052387	Trpm3	0.1196303319	0.9105345312	-0.054003125	0.7413896245	-0.191057
chr19:58296973-58301066(+)	ENSMUSG00000025089	Gfra1	-0.343493397	0.9177484643	0.4468156038	1.675207E-05	-0.786685
chr2:55427442-55436542(-)	ENSMUSG00000026824	Kcnj3	-0.580237298	0.6033988923	-0.57447561	4.282814E-08	-1.0562
chr2:66634323-66642309(+)	ENSMUSG00000075316	Scn9a	0.0687529985	0.9926893542	-0.145395638	0.2883473421	-1.12072
chr2:68470860-68477044(+)	ENSMUSG00000027030	Stk39	0.1226660881	0.9719415755	-0.084965957	0.4156602925	-1.12018
chr2:75671430-75690556(+)	ENSMUSG00000015839	Nfe2l2	0.0490989247	0.986018665	0.0838626129	0.5312041689	0.169968
chr2:109674786-109677623(-)	ENSMUSG00000048482	Bdnf	0.3293679099	0.8644903084	0.0696616305	0.8775146727	-1.0687
chr2:127481675-127485719(+)	ENSMUSG00000079056	Kcnip3	-0.515414154	0.3408899129	-0.298817895	0.0007083523	-1.07449
chr2:135894491-135898147(-)	ENSMUSG00000039943	Plcb4	0.7098198145	0.5926860334	-0.104413466	0.3379184765	-0.030264
chr2:181609802-181957990(-)	ENSMUSG00000027584	Oprl1	0.0120499059	0.9944177995	-0.505326952	4.921184E-07	0.160446
chr3:60782742-61004319(-)	ENSMUSG00000027765	P2ry1	-0.192103558	0.5060442174	-0.109144501	0.5185215467	0.749763
chr3:80800379-80851997(+)	ENSMUSG00000033981	Gria2	-1.216555107	0.5745991129	-0.49156506	5.055106E-12	-0.981679
chr3:101592328-101592581(+)	ENSMUSG00000033161	Atp1a1	-1.028248102	0.3537633784	-0.233929626	0.0531679513	-1.18998
chr4:46784711-46788504(+)	ENSMUSG00000039809	Gabbr2	1.3310521312	0.1609055133	-0.303498634	0.0164962137	-0.450315
chr4:58552293-58555304(+)	ENSMUSG00000038668	Lpar1	-0.511973278	0.7274282672	0.1851517195	0.0378675787	-0.819195
chr4:132560740-132604797(-)	ENSMUSG00000056529	Ptafr	-0.193113244	0.9673261307	0.4313546354	0.0879254679	-0.216132
chr5:15923808-15935945(-)	ENSMUSG00000040118	Cacna2d1	0.8450650334	0.1858737873	0.7358847545	2.574299E-09	-1.24297
chr5:30582864-30588672(-)	ENSMUSG00000049265	Kcnk3	0.5059393595	0.8181920646	0.1913319969	0.5451568951	-0.75535
chr5:35236149-35279078(-)	ENSMUSG00000045318	Adra2c	-0.416976374	0.0511192809	-0.368500205	0.0217133551	0.487157
chr5:43867709-43869237(-)	ENSMUSG00000029084	Cd38	0.1700508666	0.9356797221	0.3802306283	0.0994113845	-1.12311
chr5:135910531-135911651(+)	ENSMUSG00000051391	Ywhag	0.120019824	0.9465674986	-0.156355756	0.4885279085	0.444727
chr6:118163117-118170060(+)	ENSMUSG00000030110	Ret	-0.252029044	0.9492627383	-0.045756162	0.839797438	-0.98824
chr6:125241923-125242339(-)	ENSMUSG00000030337	Vamp1	-2.105684506	0.1835646872	-0.558387229	5.679823E-06	-0.713977
chr6:126636519-126640569(+)	ENSMUSG00000047976	Kcna1	-0.198656717	0.9144054087	-0.036102344	0.884347396	-0.67089
chr7:45830846-45838266(+)	ENSMUSG00000002771	Grin2d	-0.130144969	0.8742402255	-0.19475351	0.6783363649	-1.28868
chr7:51570197-51623188(-)	ENSMUSG00000030500	Slc17a6	-0.237515141	0.7052722355	0.0737330654	0.6654869991	0.0645359
chr7:57590158-57591036(-)	ENSMUSG00000033676	Gabrb3	-0.398396221	0.7794392377	-0.325063047	0.0024889939	-1.02142
chr7:75308507-75312178(+)	ENSMUSG00000053025	Sv2b	-0.366418008	0.8105102187	-0.3502195	3.549174E-06	-0.846792
chr7:91253594-91259556(-)	ENSMUSG00000052572	Dlg2	-0.314768325	0.8953219043	-0.028026406	0.9033465972	-1.2575
chr7:99263649-99269728(-)	ENSMUSG00000055407	Map6	0.1486494885	0.9105345312	-0.127707824	0.5474093107	0.035665
chr7:114635520-114636347(+)	ENSMUSG00000030669	Calca	-0.308329115	0.4775839761	-0.098255089	0.5129124813	-1.18392
chr9:96345647-96364371(+)	ENSMUSG00000032412	Atp1b3	-0.3577547	0.925482472	0.0169311681	0.9194243012	-1.00742

Table 1: All LncRNAs in mouse DRG antisense of pain genes. Log fold changes and p.values are for SNI vs Sham BALB.c strain

LncRNA genomic coordinates	ENSEMBL ID	symbol	lnc_lfc	lnc_pvalue	gene_lfc	gene_pvalue	cpc
chr1:131263308-131266930(+)	ENSMUSG00000042349	Ikbke	-1.078111727	0.397719411	0.2457525415	0.3932158227	-1.16173
chr10:39555935-39560136(-)	ENSMUSG00000019843	Fyn	-0.243497809	0.5035520725	0.043041668	0.7131938403	-0.472081
chr11:7213635-7261406(+)	ENSMUSG00000020427	Igfbp3	0.0167402906	0.989607363	0.3077295865	0.2333603316	-1.00746
chr11:55394494-55395613(+)	ENSMUSG00000018593	Sparc	0.1375173714	0.9639834139	0.2174199139	0.1764539804	-1.05264
chr11:63150892-63151309(-)	ENSMUSG00000018217	Pmp22	0.8817390532	0.4326273958	0.0345001222	0.9440613212	-0.886487
chr11:66896676-66931091(-)	ENSMUSG00000048070	Pirt	0.1938223753	0.8254926541	-0.045968414	0.7875637177	0.907072
chr11:70239888-70246096(+)	ENSMUSG00000000320	Alox12	0.0988438932	0.973077453	0.1574883068	0.8897735056	0.125631
chr11:73296005-73300736(-)	ENSMUSG00000043029	Trpv3	0.2396813077	0.7521598325	-0.156215714	0.7477299356	-0.979906
chr11:81966759-81995840(+)	ENSMUSG00000020704	Asic2	-0.351495815	0.2827416728	3.413768E-05	0.9998413947	-0.833933
chr11:83526976-83532680(+)	ENSMUSG00000035042	Ccl5	-0.522655499	0.8903792755	0.5509165533	0.5849996984	-0.685233
chr11:102739325-102762503(-)	ENSMUSG00000020926	Adam11	0.2321013698	0.730803581	-0.062294256	0.7998050125	0.539235
chr11:107935614-107937468(+)	ENSMUSG00000050965	Prkca	-0.199675752	0.9705727865	-0.02859984	0.8628297695	-0.893234
chr13:112505712-112508381(-)	ENSMUSG00000021756	Il6st	0.3297431117	0.8820463148	0.0869975911	0.2601859488	-1.01211
chr14:74638797-74642453(-)	ENSMUSG00000034997	Htr2a	-0.314947033	0.9639834139	0.055837711	0.8127659278	-1.24073
chr14:103778950-103851424(+)	ENSMUSG00000022122	Ednrb	0.2321027822	0.7655453131	-0.009434471	0.9702263737	0.298243
chr15:78081606-78108287(+)	ENSMUSG00000019146	Cacng2	0.048915626	0.9758352683	-0.305087973	0.3048619151	-0.615593
chr15:78919895-78928253(-)	ENSMUSG00000068220	Lgals1	0.222948079	0.8927790233	0.4627870665	7.638229E-10	-0.956606
chr15:79071107-79242303(-)	ENSMUSG00000068206	Pick1	0.0813402994	0.973077453	-0.027638256	0.8839176216	0.265443
chr15:101214611-101225267(-)	ENSMUSG00000000531	Grasp	-0.146283116	0.9743351145	-0.020928561	0.9628442936	-0.116895
chr15:102138265-102205121(-)	ENSMUSG00000023046	Igfbp6	-0.048028647	0.9446682396	-0.057296307	0.9303470783	0.527199
chr16:85898584-85905300(+)	ENSMUSG00000022894	Adamts5	-0.053506938	0.9705727865	0.0579465468	0.7853077524	-0.949085
chr16:87934064-87936434(+)	ENSMUSG00000022935	Grik1	-0.203429573	0.9677930437	-0.020753259	0.9361552775	-1.01687
chr16:92690953-92693516(+)	ENSMUSG00000022952	Runx1	0.1342255838	0.9705727865	0.0095130398	0.9728777749	-1.09998
chr16:94752627-94753306(+)	ENSMUSG00000043301	Kcnj6	-0.168565844	NA	0.3470856068	0.5279605773	-1.04139
chr17:86375800-86379306(-)	ENSMUSG00000045038	Prkce	-0.386538769	0.7385394353	0.0071226184	0.9769084578	-0.899925
chr18:4352995-4368040(+)	ENSMUSG00000024235	Map3k8	0.1469768665	0.969395687	0.1547033524	0.669227555	-1.1489
chr18:82405614-82406458(+)	ENSMUSG00000024553	Galr1	0.4081348311	0.785324861	0.0319199831	0.9562733313	-0.350143
chr19:6969343-6970896(+)	ENSMUSG00000024960	Plcb3	-0.715045891	0.8180007522	-0.07246381	0.7605269617	-0.934247
chr19:22435556-22448608(-)	ENSMUSG00000052387	Trpm3	-0.122783885	0.9347317631	0.010232127	0.9646533273	-0.191057
chr19:58296973-58301066(+)	ENSMUSG00000025089	Gfra1	0.1730403636	0.9732881435	0.457017995	1.317818E-06	-0.786685
chr2:55427442-55436542(-)	ENSMUSG00000026824	Kcnj3	-0.20082323	0.9387478423	-0.2065731	0.125599383	-1.0562
chr2:66634323-66642309(+)	ENSMUSG00000075316	Scn9a	-0.419457319	0.9347317631	-0.022210573	0.9319527903	-1.12072
chr2:68470860-68477044(+)	ENSMUSG00000027030	Stk39	-1.042928271	0.4961701849	-0.117325078	0.1881186421	-1.12018
chr2:75671430-75690556(+)	ENSMUSG00000015839	Nfe2l2	-0.073928483	0.9758352683	0.0389651017	0.8268121068	0.169968
chr2:109674786-109677623(-)	ENSMUSG00000048482	Bdnf	-0.23531589	0.9446610926	-0.334250402	0.2424108574	-1.0687
chr2:127481675-127485719(+)	ENSMUSG00000079056	Kcnip3	0.178893432	0.9077801239	-0.173585845	0.0845250648	-1.07449
chr2:135894491-135898147(-)	ENSMUSG00000039943	Plcb4	0.2382590536	0.9455791355	-0.077405251	0.5346915682	-0.030264
chr2:181609802-181957990(-)	ENSMUSG00000027584	Oprl1	0.1658856806	0.9639834139	-0.463996966	6.792654E-07	0.160446
chr3:60782742-61004319(-)	ENSMUSG00000027765	P2ry1	-0.04096317	0.96692948	0.0459277354	0.8411258526	0.749763
chr3:80800379-80851997(+)	ENSMUSG00000033981	Gria2	0.7836843393	0.8400002197	-0.421306441	3.234753E-10	-0.981679
chr3:101592328-101592581(+)	ENSMUSG00000033161	Atp1a1	0.4207641944	0.8608015807	-0.188540584	0.1343371599	-1.18998
chr4:46784711-46788504(+)	ENSMUSG00000039809	Gabbr2	0.6713122866	0.7497985037	-0.143473066	0.3963161582	-0.450315
chr4:58552293-58555304(+)	ENSMUSG00000038668	Lpar1	0.2387660609	0.9356855468	0.0769427646	0.5500225626	-0.819195
chr4:132560740-132604797(-)	ENSMUSG00000056529	Ptafr	-0.123339248	0.9801366717	0.3943566257	0.0981170393	-0.216132
chr5:15923808-15935945(-)	ENSMUSG00000040118	Cacna2d1	0.8156841762	0.1889214288	0.5091287054	3.342554E-05	-1.24297
chr5:30582864-30588672(-)	ENSMUSG00000049265	Kcnk3	-0.466332612	0.8912289083	0.2537765343	0.367209238	-0.75535
chr5:35236149-35279078(-)	ENSMUSG00000045318	Adra2c	-0.14461118	0.8227140887	-0.228261149	0.2277920785	0.487157
chr5:43867709-43869237(-)	ENSMUSG00000029084	Cd38	0.2385900982	0.9282341862	0.1959170178	0.5019581754	-1.12311
chr5:135910531-135911651(+)	ENSMUSG00000051391	Ywhag	0.3231610457	0.8496431381	-0.073921932	0.8061803575	0.444727
chr6:118163117-118170060(+)	ENSMUSG00000030110	Ret	-0.090794894	0.9876321853	-0.016975427	0.9552272304	-0.98824
chr6:125241923-125242339(-)	ENSMUSG00000030337	Vamp1	0.1800569004	0.9741855109	-0.481227374	4.078426E-05	-0.713977
chr6:126636519-126640569(+)	ENSMUSG00000047976	Kcna1	-0.375754473	0.8275318945	-0.22179362	0.112835847	-0.67089
chr7:45830846-45838266(+)	ENSMUSG00000002771	Grin2d	-0.311604384	0.5599885511	0.0497896167	0.945800424	-1.28868
chr7:51570197-51623188(-)	ENSMUSG00000030500	Slc17a6	-0.136785769	0.9212685871	0.0018376723	0.9930761308	0.0645359
chr7:57590158-57591036(-)	ENSMUSG00000033676	Gabrb3	-0.299312546	0.8902636021	-0.160672147	0.2380967932	-1.02142
chr7:75308507-75312178(+)	ENSMUSG00000053025	Sv2b	-0.33515086	0.8477325228	-0.288171733	7.685535E-05	-0.846792
chr7:91253594-91259556(-)	ENSMUSG00000052572	Dlg2	-0.759690261	0.6718182648	-0.049468816	0.8279471735	-1.2575
chr7:99263649-99269728(-)	ENSMUSG00000055407	Map6	0.2981164569	0.8275318945	-0.085718388	0.7403051251	0.035665
chr7:114635520-114636347(+)	ENSMUSG00000030669	Calca	-0.224843222	0.7344297262	-0.182357985	0.1064025309	-1.18392
chr9:96345647-96364371(+)	ENSMUSG00000032412	Atp1b3	0.1694189274	0.9743351145	0.0696509475	0.5935400139	-1.00742

Table 2:All LncRNAs in mouse DRG antisense of pain genes. Log fold changes and p.values are for SNI vs Sham B10.D2 strain

Appendix 5

	Dim.1	Dim.2	Dim.3
NPSI_BURNING	0.314441856707305	-0.092218177377863	-0.267987288773362
NPSI_SQUEEZING	0.296543521593314	0.140966702103109	0.281414737418969
NPSI_PRESSURE	0.334429042260101	0.345642055022025	0.152216617304376
NPSI_ELECSHOCKS	0.307958687532538	-0.430514206181964	0.14233851239341
NPSI_STABBING	0.288211204352699	-0.167543179018359	0.720355822966716
NPSI_BRUSHEVOKED	0.315262282799851	0.358482948487693	-0.077614758503885
NPSI_PRESSUREEVOKED	0.322759914289525	0.369744450168993	0.022252355946422
NPSI_COLDEVOKED	0.255132570163264	0.315931534569473	-0.365751397910472
NPSI_PANDN	0.340095191967857	-0.411221117662129	-0.297072251829438
NPSI_TINGLING	0.372987379836333	-0.323428574750983	-0.24104080625497

Table 1: Non-varimax rotated loadings for data without the spontaneous pain categorical variables. It is much more difficult to identify the highest contributing variables.

	Dim.1	Dim.2	Dim.3	Dim.4
NPSI_BURNING	0.302130884294959	-0.092173940249317	0.067460577339705	-0.235182819896692
NPSI_SQUEEZING	0.279071960504311	0.120954876609237	-0.173248916463105	0.26774479272829
NPSI_PRESSURE	0.326641767992884	0.338419682258165	-0.014151520287947	0.17123647418006
NPSI_SPONTONGOING	0.155558537325615	0.212566272710502	0.727365465007889	0.112105817881556
NPSI_ELECSHOCKS	0.293989883612732	-0.432809669597189	0.005316452827467	0.138388748663079
NPSI_STABBING	0.268144866300825	-0.190792773155487	-0.201114329371849	0.68994373116077
NPSI_SPONTPAROXYSMAL	0.269654548500442	0.018103737843322	0.526469216941567	0.009911282604728
NPSI_BRUSHEVOKED	0.2925672302872	0.340737336154331	-0.22257552676774	-0.078442017879774
NPSI_PRESSUREEVOKED	0.31084911647656	0.344260160957282	-0.126426597535849	0.007225487114492
NPSI_COLDEVOKED	0.24217065751517	0.280162399702034	-0.210739282049941	-0.43108090350353
NPSI_PANDN	0.323321630730991	-0.405789325685545	0.043493563163504	-0.281586549547539
NPSI_TINGLING	0.35141273206	-0.339609784166332	-0.081864149566949	-0.253994100655439

Table 2:Non-varimax rotated loadings for data with all variables.

Bibliography

Aboyoun, P., Pages, H., Lawrence, M., 2013. GenomicRanges: Representation and manipulation of genomic intervals.

Albrecht, A.-S., Ørom, U.A., 2016. Bidirectional expression of long ncRNA/protein-coding gene pairs in cancer. Brief. Funct. Genomics 15, 167–173. doi:10.1093/bfgp/elv048

Alexa, A., Rahnenfuhrer, J., 2010. topGO: topGO: Enrichment analysis for Gene Ontology.

Alexa, A., Rahnenfuhrer, J., Lengauer, T., 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22, 1600–1607. doi:10.1093/bioinformatics/btl140

Alschuler, K.N., Jensen, M.P., Ehde, D.M., 2012. Defining Mild, Moderate, and Severe Pain in Persons with Multiple Sclerosis. Pain Med. 13, 1358–1365. doi:10.1111/j.1526-4637.2012.01471.x

Amaral, P.P., Clark, M.B., Gascoigne, D.K., Dinger, M.E., Mattick, J.S., 2011. lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res. 39, D146–D151. doi:10.1093/nar/gkq1138

Anders, S., Huber, W., 2010. Differential expression analysis for sequence count data. Genome Biol. 11, R106. doi:10.1186/gb-2010-11-10-r106

Anders, S., Pyl, P.T., Huber, W., 2015. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169. doi:10.1093/bioinformatics/btu638

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25–29. doi:10.1038/75556

Bartus, K., Galino, J., James, N.D., Hernandez-Miranda, L.R., Dawes, J.M., Fricker, F.R., Garratt, A.N., McMahon, S.B., Ramer, M.S., Birchmeier, C., Bennett, D.L.H., Bradbury, E.J., 2016. Neuregulin-1 controls an endogenous repair mechanism after spinal cord injury. Brain 139, 1394–1416. doi:10.1093/brain/aww039

Basbaum, A.I., Bautista, D.M., Scherrer, G., Julius, D., 2009. Cellular and molecular mechanisms of pain. Cell 139, 267–284. doi:10.1016/j.cell.2009.09.028

Basso, D.M., Fisher, L.C., Anderson, A.J., Jakeman, L.B., Mctigue, D.M., Popovich, P.G., 2006. Basso Mouse Scale for Locomotion Detects Differences in Recovery after Spinal Cord Injury in Five Common Mouse Strains. J. Neurotrauma 23, 635–659. doi:10.1089/neu.2006.23.635

Bender, R., Lange, S., 2001. Adjusting for multiple testing--when and how? J. Clin. Epidemiol. 54, 343–349.

Bennett, D.L.H., Woods, C.G., 2014. Painful and painless channelopathies. Lancet Neurol. 13, 587–599. doi:10.1016/S1474-4422(14)70024-9

Bennett, G.J., Chung, J.M., Honore, M., Seltzer, Z. ’ev, 2003. Models of neuropathic pain in the rat. Curr. Protoc. Neurosci. Editor. Board Jacqueline N Crawley Al Chapter 9, Unit 9.14. doi:10.1002/0471142301.ns0914s22

Bouhassira, D., Attal, N., Alchaar, H., Boureau, F., Brochet, B., Bruxelle, J., Cunin, G., Fermanian, J., Ginies, P., Grun-Overdyking, A., Jafari-Schluep, H., Lantéri-Minet, M., Laurent, B., Mick, G., Serrie, A., Valade, D., Vicaut, E., 2005. Comparison of pain syndromes associated with nervous or somatic lesions and development of a new neuropathic pain diagnostic questionnaire (DN4). Pain 114, 29–36. doi:10.1016/j.pain.2004.12.010

Bouhassira, D., Attal, N., Fermanian, J., Alchaar, H., Gautron, M., Masquelier, E., Rostaing, S., Lanteri-Minet, M., Collin, E., Grisart, J., Boureau, F., 2004. Development and validation of the Neuropathic Pain Symptom Inventory. Pain 108, 248–257. doi:10.1016/j.pain.2003.12.024

Bridges, D., Ahmad, K., Rice, A.S., 2001. The synthetic cannabinoid WIN55,212-2 attenuates hyperalgesia and allodynia in a rat model of neuropathic pain. Br. J. Pharmacol. 133, 586–594. doi:10.1038/sj.bjp.0704110

Bril, V., Perkins, B.A., 2002. Validation of the Toronto Clinical Scoring System for diabetic polyneuropathy. Diabetes Care 25, 2048–2052.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., Rinn, J.L., 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 25, 1915–1927. doi:10.1101/gad.17446611

Calvo, M., Dawes, J.M., Bennett, D.L., 2012. The role of the immune system in the generation of neuropathic pain. Lancet Neurol. 11, 629–642. doi:10.1016/S1474-4422(12)70134-5

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421. doi:10.1186/1471-2105-10-421

Casella, G., Garzetti, L., Gatta, A.T., Finardi, A., Maiorino, C., Ruffini, F., Martino, G., Muzio, L., Furlan, R., 2016. IL4 induces IL6-producing M2 macrophages associated to inhibition of neuroinflammation in vitro and in vivo. J. Neuroinflammation 13. doi:10.1186/s12974-016-0596-5

Cheng, H.T., Dauch, J.R., Porzio, M.T., Yanik, B.M., Hsieh, W., Smith, A.G., Singleton, J.R., Feldman, E.L., 2013. Increased Axonal Regeneration and Swellings in Intraepidermal Nerve Fibers Characterize Painful Phenotypes of Diabetic Neuropathy. J. Pain 14, 941–947. doi:10.1016/j.jpain.2013.03.005

Chinzei, N., Hashimoto, S., Fujishiro, T., Hayashi, S., Kanzaki, N., Uchida, S., Kuroda, R., Kurosaka, M., 2016. Inflammation and Degeneration in Cartilage Samples from Patients with Femoroacetabular Impingement. J. Bone Jt. Surg. 98, 135–141. doi:10.2106/JBJS.O.00443

Costigan, M., Befort, K., Karchewski, L., Griffin, R.S., D'Urso, D., Allchorne, A., Sitarski, J., Mannion, J.W., Pratt, R.E., Woolf, C.J., 2002. Replicate high-density rat genome oligonucleotide microarrays reveal hundreds of regulated genes in the dorsal root ganglion after peripheral nerve injury. BMC Neurosci. 3, 16.

Costigan, M., Moss, A., Latremoliere, A., Johnston, C., Verma-Gandhu, M., Herbert, T.A., Barrett, L., Brenner, G.J., Vardeh, D., Woolf, C.J., Fitzgerald, M., 2009. T-Cell Infiltration and Signaling in the Adult Dorsal Spinal Cord Is a Major Contributor to Neuropathic Pain-Like Hypersensitivity. J. Neurosci. 29, 14415–14422. doi:10.1523/JNEUROSCI.4569-09.2009

Costigan, M., Scholz, J., Woolf, C.J., 2009. Neuropathic Pain: A Maladaptive Response of the Nervous System to Damage. Annu. Rev. Neurosci. 32, 1–32. doi:10.1146/annurev.neuro.051508.135531

Cruccu, G., Truini, A., 2009. Tools for Assessing Neuropathic Pain. PLoS Med. 6. doi:10.1371/journal.pmed.1000045

Dawes, J.M., Antunes-Martins, A., Perkins, J.R., Paterson, K.J., Sisignano, M., Schmid, R., Rust, W., Hildebrandt, T., Geisslinger, G., Orengo, C., Bennett, D.L., McMahon, S.B., 2014. Genome-wide transcriptional profiling of skin and dorsal root ganglia after ultraviolet-B-induced inflammation. PloS One 9, e93338. doi:10.1371/journal.pone.0093338

Decosterd, I., Woolf, C.J., 2000. Spared nerve injury: an animal model of persistent peripheral neuropathic pain. Pain 87, 149–158.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. doi:10.1093/bioinformatics/bts635

Dowle, M., Srinivasan, A., Short, T., Saporta, S.L. with contributions from R., Antonyan, E., 2015. data.table: Extension of Data.frame.

Durinck, S., Spellman, P.T., Birney, E., Huber, W., 2009. Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor package biomaRt. Nat. Protoc. 4, 1184–1191. doi:10.1038/nprot.2009.97

Eppig, J.T., Blake, J.A., Bult, C.J., Kadin, J.A., Richardson, J.E., The Mouse Genome Database Group, 2015. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. Nucleic Acids Res. 43, D726–D736. doi:10.1093/nar/gku967

Estacion, M., Dib-Hajj, S.D., Benke, P.J., te Morsche, R.H.M., Eastman, E.M., Macala, L.J., Drenth, J.P.H., Waxman, S.G., 2008. NaV1.7 Gain-of-Function Mutations as a Continuum: A1632E Displays Physiological Changes Associated with Erythromelalgia and Paroxysmal Extreme Pain Disorder Mutations and Produces Symptoms of Both Disorders. J. Neurosci. 28, 11079–11088. doi:10.1523/JNEUROSCI.3443-08.2008

Ewing, B., Green, P., 1998. Base-Calling of Automated Sequencer Traces Using *Phred.* II. Error Probabilities. Genome Res. 8, 186–194. doi:10.1101/gr.8.3.186

FactoMineR: An R Package for Multivariate Analysis | Lê | Journal of Statistical Software [WWW Document], n.d. URL https://www.jstatsoft.org/article/view/v025i01 (accessed 5.3.16).

Fertleman, C.R., Baker, M.D., Parker, K.A., Moffatt, S., Elmslie, F.V., Abrahamsen, B., Ostman, J., Klugbauer, N., Wood, J.N., Gardiner, R.M., Rees, M., 2006. SCN9A mutations in paroxysmal extreme pain disorder: allelic variants underlie distinct channel defects and phenotypes. Neuron 52, 767–774. doi:10.1016/j.neuron.2006.10.006

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kahari, A.K., Keefe, D., Keenan, S., Kinsella, R., Komorowska, M., Koscielny, G., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Muffato, M., Overduin, B., Pignatelli, M., Pritchard, B., Riat, H.S., Ritchie, G.R.S., Ruffier, M., Schuster, M., Sobral, D., Tang, Y.A., Taylor, K., Trevanion, S., Vandrovcova, J., White, S., Wilson, M., Wilder, S.P., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernandez-Suarez, X.M., Harrow, J., Herrero, J., Hubbard, T.J.P., Parker, A., Proctor, G., Spudich, G., Vogel, J., Yates, A., Zadissa, A., Searle, S.M.J., 2011. Ensembl 2012. Nucleic Acids Res. 40, D84–D90. doi:10.1093/nar/gkr991

Freeman, R., Baron, R., Bouhassira, D., Cabrera, J., Emir, B., 2014. Sensory profiles of patients with neuropathic pain based on the neuropathic pain symptoms and signs. Pain 155, 367–376. doi:10.1016/j.pain.2013.10.023

Galili, T., 2015. dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. Bioinformatics. doi:10.1093/bioinformatics/btv428

Gawad, C., Koh, W., Quake, S.R., 2016. Single-cell genome sequencing: current state of the science. Nat. Rev. Genet. 17, 175–188. doi:10.1038/nrg.2015.16

Gentleman, R.C., Carey, V.J., Bates, D.M., others, 2004. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol. 5, R80.

Gold, M.S., Gebhart, G.F., 2010. Nociceptor sensitization in pain pathogenesis. Nat. Med. 16, 1248–1257. doi:10.1038/nm.2235

Gong, L., Wu, J., Zhou, S., Wang, Y., Qin, J., Yu, B., Gu, X., Yao, C., 2016. Global analysis of transcriptome in dorsal root ganglia following peripheral nerve injury in rats. Biochem. Biophys. Res. Commun. 478, 206–212. doi:10.1016/j.bbrc.2016.07.067

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., Cabili, M.N., Jaenisch, R., Mikkelsen, T.S., Jacks, T., Hacohen, N., Bernstein, B.E., Kellis, M., Regev, A., Rinn, J.L., Lander, E.S., 2009. Chromatin signature reveals over a thousand highly conserved large

non-coding RNAs in mammals. Nature 458, 223–227. doi:10.1038/nature07672
Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., Rinn, J.L., Lander, E.S., Regev, A., 2010. Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat. Biotechnol. 28, 503–510. doi:10.1038/nbt.1633
Han, T.W., Jan, L.Y., 2013. Making antisense of pain. Nat. Neurosci. 16, 986–987. doi:10.1038/nn.3475
Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J.M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R., Hubbard, T.J., 2012. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22, 1760–1774. doi:10.1101/gr.135350.111
Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: A K-Means Clustering Algorithm. Appl. Stat. 28, 100. doi:10.2307/2346830
Hennig, C., 2015. fpc: Flexible Procedures for Clustering.
Husson, F., Josse, J., 2015. missMDA: Handling Missing Values with Multivariate Data Analysis.
Husson, F., Josse, J., Le, S., Mazet, J., 2016. FactoMineR: Multivariate Exploratory Data Analysis and Data Mining.
Ilott, N.E., Ponting, C.P., 2013. Predicting long non-coding RNAs using RNA sequencing. Methods San Diego Calif 63, 50–59. doi:10.1016/j.ymeth.2013.03.019
Jaggi, A.S., Jain, V., Singh, N., 2011. Animal models of neuropathic pain. Fundam. Clin. Pharmacol. 25, 1–28. doi:10.1111/j.1472-8206.2009.00801.x
JAX® Mice & Services [WWW Document], n.d. URL https://www.jax.org/jax-mice-and-services (accessed 12.1.15).
Jiang, B.-C., Sun, W.-X., He, L.-N., Cao, D.-L., Zhang, Z.-J., Gao, Y.-J., 2015. Identification of lncRNA expression profile in the spinal cord of mice following spinal nerve ligation-induced neuropathic pain. Mol. Pain 11, 43. doi:10.1186/s12990-015-0047-9
Josse, J., Husson, F., 2012. Handling missing values in exploratory multivariate data analysis methods. J. Société Fr. Stat. 153, 79–99.
Jr, F.E.H., Dupont, with contributions from C., others, many, 2016. Hmisc: Harrell Miscellaneous.
Kaiser, H.F., 1958. The varimax criterion for analytic rotation in factor analysis. Psychometrika 23, 187–200. doi:10.1007/BF02289233

Kapusta, A., Feschotte, C., 2014. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. Trends Genet. 30, 439–452. doi:10.1016/j.tig.2014.08.004
Kent, W.J., 2002. BLAT—The BLAST-Like Alignment Tool. Genome Res. 12, 656–664. doi:10.1101/gr.229202
Khan, H., Khan, H., Sherwani, S., Ekhzaimy, A., Masood, A., Sakharkar, M., 2016. Significance of HbA1c Test in Diagnosis and Prognosis of Diabetic Patients. Biomark. Insights 95. doi:10.4137/BMI.S38440
Khuong, T.M., Neely, G.G., 2013. Conserved systems and functional genomic assessment of nociception. FEBS J. 280, 5298–5306. doi:10.1111/febs.12464
Kiguchi, N., Kobayashi, Y., Saika, F., Sakaguchi, H., Maeda, T., Kishioka, S., 2015. Peripheral interleukin-4 ameliorates inflammatory macrophage-dependent neuropathic pain: PAIN 156, 684–693. doi:10.1097/j.pain.0000000000000097
Koenig, J., Werdehausen, R., Linley, J.E., Habib, A.M., Vernon, J., Lolignier, S., Eijkelkamp, N., Zhao, J., Okorokov, A.L., Woods, C.G., Wood, J.N., Cox, J.J., 2015. Regulation of Nav1.7: A Conserved SCN9A Natural Antisense Transcript Expressed in Dorsal Root Ganglia. PloS One 10, e0128830. doi:10.1371/journal.pone.0128830
Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., Gao, G., 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. 35, W345-349. doi:10.1093/nar/gkm391
LaCroix-Fralish, M.L., Austin, J.-S., Zheng, F.Y., Levitin, D.J., Mogil, J.S., 2011. Patterns of pain: meta-analysis of microarray studies of pain. Pain 152, 1888–1898. doi:10.1016/j.pain.2011.04.014
Lacroix-Fralish, M.L., Ledoux, J.B., Mogil, J.S., 2007. The Pain Genes Database: An interactive web browser of pain-related transgenic knockout studies. Pain 131, 3.e1-4. doi:10.1016/j.pain.2007.04.041
Lauria, G., Hsieh, S.T., Johansson, O., Kennedy, W.R., Leger, J.M., Mellgren, S.I., Nolano, M., Merkies, I.S.J., Polydefkis, M., Smith, A.G., Sommer, C., Valls-Solé, J., 2010. European Federation of Neurological Societies/Peripheral Nerve Society Guideline on the use of skin biopsy in the diagnosis of small fiber neuropathy. Report of a joint task force of the European Fe-deration of Neurological Societies and the Peripheral Ne: EFNS/PNS guideline on skin biopsy. Eur. J. Neurol. 17, 903-e49. doi:10.1111/j.1468-1331.2010.03023.x
Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., Carey, V.J., 2013. Software for Computing and Annotating Genomic Ranges. PLOS Comput Biol 9, e1003118. doi:10.1371/journal.pcbi.1003118
Lötsch, J., Doehring, A., Mogil, J.S., Arndt, T., Geisslinger, G., Ultsch, A., 2013. Functional genomics of pain in analgesic drug development and therapy. Pharmacol. Ther. 139, 60–70. doi:10.1016/j.pharmthera.2013.04.004

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550. doi:10.1186/s13059-014-0550-8
Lu, B., Zhang, Q., Wang, H., Wang, Y., Nakayama, M., Ren, D., 2010. Extracellular Calcium Controls Background Current and Neuronal Excitability via an UNC79-UNC80-NALCN Cation Channel Complex. Neuron 68, 488–499. doi:10.1016/j.neuron.2010.09.014
Ma, C.H.E., Omura, T., Cobos, E.J., Latrémolière, A., Ghasemlou, N., Brenner, G.J., van Veen, E., Barrett, L., Sawada, T., Gao, F., Coppola, G., Gertler, F., Costigan, M., Geschwind, D., Woolf, C.J., 2011. Accelerating axonal growth promotes motor recovery after peripheral nerve injury in mice. J. Clin. Invest. 121, 4332–4347. doi:10.1172/JCI58675
Maden, C.H., Gomes, J., Schwarz, Q., Davidson, K., Tinker, A., Ruhrberg, C., 2012. NRP1 and NRP2 cooperate to regulate gangliogenesis, axon guidance and target innervation in the sympathetic nervous system. Dev. Biol. 369, 277–285. doi:10.1016/j.ydbio.2012.06.026
Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., 2015. cluster: Cluster Analysis Basics and Extensions.
Marques, A.C., Ponting, C.P., 2014. Intergenic lncRNAs and the evolution of gene expression. Curr. Opin. Genet. Dev. 27, 48–53. doi:10.1016/j.gde.2014.03.009
Martin, B., Marchaland, C., Phillips, J., Chapouthier, G., Spach, C., Motta, R., 1992. Recombinant congenic strains of mice from B10.D2 and DBA/2: Their contribution to behavior genetic research and application to audiogenic seizures. Behav. Genet. 22, 685–701. doi:10.1007/BF01066639
Mechenthaler, I., 2008. Galanin – 25 years with a multitalented neuropeptide: Galanin and the neuroendocrine axes. Cell. Mol. Life Sci. 65, 1826–1835. doi:10.1007/s00018-008-8157-4
Merskey, H., 1968. Psychological aspects of pain. Postgrad. Med. J. 44, 297–306.
Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B., Raney, B.J., Pohl, A., Malladi, V.S., Li, C.H., Lee, B.T., Learned, K., Kirkup, V., Hsu, F., Heitner, S., Harte, R.A., Haeussler, M., Guruvadoo, L., Goldman, M., Giardine, B.M., Fujita, P.A., Dreszer, T.R., Diekhans, M., Cline, M.S., Clawson, H., Barber, G.P., Haussler, D., Kent, W.J., 2012. The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res. 41, D64–D69. doi:10.1093/nar/gks1048
Minett, M.S., Pereira, V., Sikandar, S., Matsuyama, A., Lolignier, S., Kanellopoulos, A.H., Mancini, F., Iannetti, G.D., Bogdanov, Y.D., Santana-Varela, S., Millet, Q., Baskozos, G., MacAllister, R., Cox, J.J., Zhao, J., Wood, J.N., 2015. Endogenous opioids contribute to insensitivity to pain in humans and mice lacking sodium channel Nav1.7. Nat. Commun. 6, 8967. doi:10.1038/ncomms9967
Miró, J., de la Vega, R., Solé, E., Racine, M., Jensen, M.P., Gálan, S., Engel, J.M., 2016. Defining mild, moderate, and severe pain in young

people with physical disabilities. Disabil. Rehabil. 1–8. doi:10.1080/09638288.2016.1185469
Mogil, J.S., 2009. Animal models of pain: progress and challenges. Nat. Rev. Neurosci. 10, 283–294. doi:10.1038/nrn2606
Mogil, J.S., Adhikari, S.M., 1999. Hot and cold nociception are genetically correlated. J. Neurosci. Off. J. Soc. Neurosci. 19, RC25.
Mogil, J.S., Davis, K.D., Derbyshire, S.W., 2010. The necessity of animal models in pain research: Pain 151, 12–17. doi:10.1016/j.pain.2010.07.015
Mogil, J.S., Richards, S.P., O'Toole, L.A., Helms, M.L., Mitchell, S.R., Belknap, J.K., 1997. Genetic sensitivity to hot-plate nociception in DBA/2J and C57BL/6J inbred mouse strains: possible sex-specific mediation by delta2-opioid receptors. Pain 70, 267–277.
Neely, G.G., Hess, A., Costigan, M., Keene, A.C., Goulas, S., Langeslag, M., Griffin, R.S., Belfer, I., Dai, F., Smith, S.B., Diatchenko, L., Gupta, V., Xia, C., Amann, S., Kreitz, S., Heindl-Erdmann, C., Wolz, S., Ly, C.V., Arora, S., Sarangi, R., Dan, D., Novatchkova, M., Rosenzweig, M., Gibson, D.G., Truong, D., Schramek, D., Zoranovic, T., Cronin, S.J.F., Angjeli, B., Brune, K., Dietzl, G., Maixner, W., Meixner, A., Thomas, W., Pospisilik, J.A., Alenius, M., Kress, M., Subramaniam, S., Garrity, P.A., Bellen, H.J., Woolf, C.J., Penninger, J.M., 2010. A Genome-wide Drosophila Screen for Heat Nociception Identifies α2δ3 as an Evolutionarily Conserved Pain Gene. Cell 143, 628–638. doi:10.1016/j.cell.2010.09.047
Nuzzo, R., 2014. Scientific method: Statistical errors. Nature 506, 150–152. doi:10.1038/506150a
Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., Yamanaka, I., Kiyosawa, H., Yagi, K., Tomaru, Y., Hasegawa, Y., Nogami, A., Schönbach, C., Gojobori, T., Baldarelli, R., Hill, D.P., Bult, C., Hume, D.A., Quackenbush, J., Schriml, L.M., Kanapin, A., Matsuda, H., Batalov, S., Beisel, K.W., Blake, J.A., Bradt, D., Brusic, V., Chothia, C., Corbani, L.E., Cousins, S., Dalla, E., Dragani, T.A., Fletcher, C.F., Forrest, A., Frazer, K.S., Gaasterland, T., Gariboldi, M., Gissi, C., Godzik, A., Gough, J., Grimmond, S., Gustincich, S., Hirokawa, N., Jackson, I.J., Jarvis, E.D., Kanai, A., Kawaji, H., Kawasawa, Y., Kedzierski, R.M., King, B.L., Konagaya, A., Kurochkin, I.V., Lee, Y., Lenhard, B., Lyons, P.A., Maglott, D.R., Maltais, L., Marchionni, L., McKenzie, L., Miki, H., Nagashima, T., Numata, K., Okido, T., Pavan, W.J., Pertea, G., Pesole, G., Petrovsky, N., Pillai, R., Pontius, J.U., Qi, D., Ramachandran, S., Ravasi, T., Reed, J.C., Reed, D.J., Reid, J., Ring, B.Z., Ringwald, M., Sandelin, A., Schneider, C., Semple, C. a. M., Setou, M., Shimada, K., Sultana, R., Takenaka, Y., Taylor, M.S., Teasdale, R.D., Tomita, M., Verardo, R., Wagner, L., Wahlestedt, C., Wang, Y., Watanabe, Y., Wells, C., Wilming, L.G., Wynshaw-Boris, A., Yanagisawa, M., Yang, I., Yang, L., Yuan, Z., Zavolan, M., Zhu, Y., Zimmer, A., Carninci, P., Hayatsu, N., Hirozane-Kishikawa, T.,

Konno, H., Nakamura, M., Sakazume, N., Sato, K., Shiraki, T., Waki, K., Kawai, J., Aizawa, K., Arakawa, T., Fukuda, S., Hara, A., Hashizume, W., Imotani, K., Ishii, Y., Itoh, M., Kagawa, I., Miyazaki, A., Sakai, K., Sasaki, D., Shibata, K., Shinagawa, A., Yasunishi, A., Yoshino, M., Waterston, R., Lander, E.S., Rogers, J., Birney, E., Hayashizaki, Y., 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420, 563–573. doi:10.1038/nature01266

Ostrakhovitch, E.A., Semenikhin, O.A., 2011. p53-mediated regulation of neuronal differentiation via regulation of dual oxidase maturation factor 1. Neurosci. Lett. 494, 80–85. doi:10.1016/j.neulet.2011.02.061

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., Hellmann, I., 2016. The impact of amplification on differential expression analyses by RNA-seq. Sci. Rep. 6, 25533. doi:10.1038/srep25533

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., Soldatov, A., 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. 37, e123–e123. doi:10.1093/nar/gkp596

Perkins, J.R., 2013. Functional genomics and bioinformatics protocols for the elucidation of pain (Doctoral). UCL (University College London).

Perkins, J.R., Antunes-Martins, A., Calvo, M., Grist, J., Rust, W., Schmid, R., Hildebrandt, T., Kohl, M., Orengo, C., McMahon, S.B., Bennett, D.L., 2014. A comparison of RNA-seq and exon arrays for whole genome transcription profiling of the L5 spinal nerve transection model of neuropathic pain in the rat. Mol. Pain 10, 7. doi:10.1186/1744-8069-10-7

Perkins, J.R., Lees, J., Antunes-Martins, A., Diboun, I., McMahon, S.B., Bennett, D.L.H., Orengo, C., 2013. PainNetworks: a web-based resource for the visualisation of pain-related genes in the context of their network associations. Pain 154, 2586.e1-12. doi:10.1016/j.pain.2013.09.003

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., Salzberg, S.L., 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33, 290–295. doi:10.1038/nbt.3122

Ponting, C.P., Oliver, P.L., Reik, W., 2009. Evolution and Functions of Long Noncoding RNAs. Cell 136, 629–641. doi:10.1016/j.cell.2009.02.006

Principal Component Analysis, 2002. , Springer Series in Statistics. Springer-Verlag, New York.

Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., Murphy, M.R., O'Leary, N.A., Pujar, S., Rajput, B., Rangwala, S.H., Riddick, L.D., Shkeda, A., Sun, H., Tamez, P., Tully, R.E., Wallin, C., Webb, D., Weber, J., Wu, W., DiCuccio, M., Kitts, P., Maglott, D.R., Murphy, T.D., Ostell, J.M., 2014. RefSeq: an

update on mammalian reference sequences. Nucleic Acids Res. 42, D756-763. doi:10.1093/nar/gkt1114

R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Ranade, S.S., Woo, S.-H., Dubin, A.E., Moshourab, R.A., Wetzel, C., Petrus, M., Mathur, J., Bégay, V., Coste, B., Mainquist, J., Wilson, A.J., Francisco, A.G., Reddy, K., Qiu, Z., Wood, J.N., Lewin, G.R., Patapoutian, A., 2014. Piezo2 is the major transducer of mechanical forces for touch sensation in mice. Nature 516, 121–125. doi:10.1038/nature13980

Ren, D., 2011. Sodium Leak Channels in Neuronal Excitability and Rhythmic Behaviors. Neuron 72, 899–911. doi:10.1016/j.neuron.2011.12.007

Rigaud, M., Gemes, G., Barabas, M.-E., Chernoff, D.I., Abram, S.E., Stucky, C.L., Hogan, Q.H., 2008. Species and strain differences in rodent sciatic nerve anatomy: implications for studies of neuropathic pain. Pain 136, 188–201. doi:10.1016/j.pain.2008.01.016

Rolke, R., Baron, R., Maier, C., Tölle, T.R., Treede, R.-D., Beyer, A., Binder, A., Birbaumer, N., Birklein, F., Bötefür, I.C., Braune, S., Flor, H., Huge, V., Klug, R., Landwehrmeyer, G.B., Magerl, W., Maihöfner, C., Rolko, C., Schaub, C., Scherens, A., Sprenger, T., Valet, M., Wasserka, B., 2006. Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): standardized protocol and reference values. Pain 123, 231–243. doi:10.1016/j.pain.2006.01.041

Seqc/Maqc-Iii Consortium, 2014. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat. Biotechnol. advance online publication. doi:10.1038/nbt.2957

Shields, S.D., Eckert, W.A., Basbaum, A.I., 2003. Spared nerve injury model of neuropathic pain in the mouse: a behavioral and anatomic analysis. J. Pain Off. J. Am. Pain Soc. 4, 465–470.

Smith, S.B., Marker, C.L., Perry, C., Liao, G., Sotocinal, S.G., Austin, J.-S., Melmed, K., David Clark, J., Peltz, G., Wickman, K., Mogil, J.S., 2008. Quantitative trait locus and computational mapping identifies Kcnj9 (GIRK3) as a candidate gene affecting analgesia from multiple drug classes: Pharmacogenet. Genomics 18, 231–241. doi:10.1097/FPC.0b013e3282f55ab2

Sorge, R.E., Trang, T., Dorfman, R., Smith, S.B., Beggs, S., Ritchie, J., Austin, J.-S., Zaykin, D.V., Meulen, H.V., Costigan, M., Herbert, T.A., Yarkoni-Abitbul, M., Tichauer, D., Livneh, J., Gershon, E., Zheng, M., Tan, K., John, S.L., Slade, G.D., Jordan, J., Woolf, C.J., Peltz, G., Maixner, W., Diatchenko, L., Seltzer, Z. ’ev, Salter, M.W., Mogil, J.S., 2012. Genetically determined P2X7 receptor pore formation regulates variability in chronic pain sensitivity. Nat. Med. 18, 595–599. doi:10.1038/nm.2710

Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., Wu, C.H., 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23, 1282–1288. doi:10.1093/bioinformatics/btm098

Suzuki, R., Shimodaira, H., 2015. pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling.

Tedeschi, A., Omura, T., Costigan, M., 2016. CNS repair and axon regeneration: Using genetic variation to determine mechanisms. Exp. Neurol. doi:10.1016/j.expneurol.2016.05.004

Themistocleous, A.C., Ramirez, J.D., Shillo, P.R., Lees, J.G., Selvarajah, D., Orengo, C., Tesfaye, S., Rice, A.S.C., Bennett, D.L.H., 2016. The Pain in Neuropathy Study (PiNS): a cross-sectional observational study determining the somatosensory phenotype of painful and painless diabetic neuropathy. PAIN 157, 1132–1145. doi:10.1097/j.pain.0000000000000491

Thierry-Mieg, D., Thierry-Mieg, J., 2006. AceView: a comprehensive cDNA-supported gene and transcripts annotation. Genome Biol. 7, 1–14. doi:10.1186/gb-2006-7-s1-s12

Thygesen, H.H., Zwinderman, A.H., 2005. Modelling the correlation between the activities of adjacent genes in drosophila. BMC Bioinformatics 6, 10. doi:10.1186/1471-2105-6-10

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7, 562–578. doi:10.1038/nprot.2012.016

Treede, R.-D., Jensen, T.S., Campbell, J.N., Cruccu, G., Dostrovsky, J.O., Griffin, J.W., Hansson, P., Hughes, R., Nurmikko, T., Serra, J., 2008. Neuropathic pain Redefinition and a grading system for clinical and research purposes. Neurology 70, 1630–1635. doi:10.1212/01.wnl.0000282763.29778.59

Ulitsky, I., Bartel, D.P., 2013. lincRNAs: Genomics, Evolution, and Mechanisms. Cell 154, 26–46. doi:10.1016/j.cell.2013.06.020

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63. doi:10.1038/nrg2484

Ward, J.H., 1963. Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. 58, 236–244. doi:10.1080/01621459.1963.10500845

White, F.A., Bhangoo, S.K., Miller, R.J., 2005. Chemokines: Integrators of Pain and Inflammation. Nat. Rev. Drug Discov. 4, 834–844. doi:10.1038/nrd1852

White, F.A., Wilson, N., 2008. Chemokines as Pain Mediators and Modulators. Curr. Opin. Anaesthesiol. 21, 580–585. doi:10.1097/ACO.0b013e32830eb69d

Woolf, C.J., Salter, M.W., 2000. Neuronal Plasticity: Increasing the Gain in Pain. Science 288, 1765–1768. doi:10.1126/science.288.5472.1765

Wu, S., Marie Lutz, B., Miao, X., Liang, L., Mo, K., Chang, Y.-J., Du, P., Soteropoulos, P., Tian, B., Kaufman, A.G., Bekker, A., Hu, Y., Tao, Y.-X., 2016. Dorsal root ganglion transcriptome analysis following peripheral nerve injury in mice. Mol. Pain 12. doi:10.1177/1744806916629048

Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., Zhu, W., Wu, W., Chen, R., Zhao, Y., 2014. NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res. 42, D98-103. doi:10.1093/nar/gkt1222

Xu, J., Bai, J., Zhang, X., Lv, Y., Gong, Y., Liu, L., Zhao, H., Yu, F., Ping, Y., Zhang, G., Lan, Y., Xiao, Y., Li, X., 2016. A comprehensive overview of lncRNA annotation resources. Brief. Bioinform. doi:10.1093/bib/bbw015

Yamanaka, H., Kobayashi, K., Okubo, M., Fukuoka, T., Noguchi, K., 2011. Increase of close homolog of cell adhesion molecule L1 in primary afferent by nerve injury and the contribution to neuropathic pain. J. Comp. Neurol. 519, 1597–1615. doi:10.1002/cne.22588

Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Girón, C.G., Gordon, L., Hourlier, T., Hunt, S.E., Janacek, S.H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F.J., Maurel, T., McLaren, W., Murphy, D.N., Nag, R., Nuhn, M., Parker, A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H.S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S.P., Zadissa, A., Birney, E., Harrow, J., Muffato, M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S.J., Cunningham, F., Aken, B.L., Zerbino, D.R., Flicek, P., 2016. Ensembl 2016. Nucleic Acids Res. 44, D710–D716. doi:10.1093/nar/gkv1157

Young, R.S., Marques, A.C., Tibbit, C., Haerty, W., Bassett, A.R., Liu, J.-L., Ponting, C.P., 2012. Identification and Properties of 1,119 Candidate LincRNA Loci in the Drosophila melanogaster Genome. Genome Biol. Evol. 4, 427–442. doi:10.1093/gbe/evs020

Zhao, X., Tang, Z., Zhang, H., Atianjoh, F.E., Zhao, J.-Y., Liang, L., Wang, W., Guan, X., Kao, S.-C., Tiwari, V., Gao, Y.-J., Hoffman, P.N., Cui, H., Li, M., Dong, X., Tao, Y.-X., 2013. A long noncoding RNA contributes to neuropathic pain by silencing Kcna2 in primary afferent neurons. Nat. Neurosci. 16, 1024–1031. doi:10.1038/nn.3438